Copyright Statement

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand).

This thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognize the author's right to be identified as the author of this thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from their thesis.

General copyright and disclaimer

In addition to the above conditions, authors give their consent for the digital copy of their work to be used subject to the conditions specified on the Library Thesis Consent Form and Deposit Licence.
Optimal Allocation of Intensive Care Unit Nurses to Patient-At-Risk-Team

Ali Haji Vahabzadeh

Abstract

Medical studies report an association between critical care nursing and patient mortality in an Intensive Care Unit (ICU). Increasing the nurse-to-patient ratio in hospitals is not feasible, mainly due to an ever-growing nursing shortage and limited hospital budget. However, an alternative approach could be reducing the ICU demands through preventing unnecessary and unplanned admissions from wards. This is the primary goal of a nurse-led Patient-AT-Risk-Team (PART) capable of identifying deteriorating patients in wards who could be provided with the required services outside of ICU.

Specifically, PART is a nurse-led outreach team with the following aims: First, to avert ICU admission by identifying patients who are deteriorating; second, to safely discharge from ICU by following up patients discharged to the ward; and, finally, to prevent admissions of patients who are either not in critical condition despite showing signs or can be treated in the ward. Constructing such a costly team though is often justifiable when it is staffed with the same ICU nurses, but, lowering the level of ICU nursing could lead to worse patient outcomes such as mortality. Thus, we aim to investigate what nurse allocation policy between PART and ICU would result in best possible outcomes for both patients and hospitals.

This thesis is comprised of two distinct but supplementary research papers. Paper I provides econometric models to estimate the impact of the occupancy level of critical care nurses in both PART and ICU on patient hospital length of stay. Paper II proposes queueing and simulation models to obtain the optimal nurse allocation policy between PART and ICU aimed at minimising the ICU mortality rate. The proposed models were validated at
Middlemore Hospital using a 12-month history (1 July 2015 to 30 June 2016) of 8,576 visits of the PART to 2,662 patients.

The findings of paper I indicate that when the ICU nurses are busy, length of stay increases by 3% (nearly 125 patient-days), and the high utilisation rate of the PART nurses increases the length of stay by 2% (almost 70 patient-days). The results would seem to suggest roughly $300,000 annual saving by allocating a new nurse to PART per shift. Simulation results in paper II suggest that establishing PART at hospitals even with one ICU nurse might reduce ICU deaths significantly (almost 35%). Lastly, the evidence from the “what-if” analysis implies that, in an ICU with 18 beds, the configuration of 3 PART and 15 ICU nurses minimises the ICU mortality rate.
Acknowledgements

Firstly, I would like to express my special appreciation and thanks to my supervisor, Doctor Valery Pavlov for his continuous and tremendous supports over my PhD journey. I specially thank him for his patience, dedicated help, inspiration and encouragement throughout my PhD study. Without his constructive support and guidance, this thesis would not exist.

I would like to thank my co-supervisor, Professor Tava Olsen for her insightful comments in the first year of my PhD study. I thank colleagues in the Department of Information Systems and Operations Management, especially Associate Professor Tiru Arthanari, Dr Gabrielle Mary Peko and Dr Timofey Shalpegin for providing me with valuable comments and feedback. My sincere thanks goes to ex-Head of Department, Professor Michael Myers for creating a positive and competitive atmosphere for all the staff at the department. My words of thanks should also go to Josephine Lee, the Professional Teaching Fellow and Chrissy Bretherton, the Group Services Team Leader at the department for their full support and guidance.

I am hugely appreciative to the ICU consultant at the Middlemore Hospital, Dr Alex Kazemi for all his valuable time, advice and support provided during this research. I especially thank him and the PART nurses at Middlemore Hospital for supplying the required data and information to validate our proposed models.

As an international student I would not have been able to fulfil my dream without financial support from the University of Auckland. I must thank the scholarship committee at the University of Auckland for the Doctoral Scholarship award.

Special mention goes to my friends at the University of Auckland who have made my
PhD life very pleasant and memorable. Thanks to Foad Marzoughi, Hamid Shahbazzadeh, Farzan Kolini, Mojtaba Mahdavi, Sina Mashinchi, Dusan Ku, Xiaojing Liu and Shandong Mou for all the joys and energy that they shared with me.

I owe my deepest gratitude towards my family, especially my mother, who has devoted all her life to my success. Her love and support has always been my strength. Her patience and sacrifice will remain my inspiration throughout my life. I also thank my sisters and father for supporting me spiritually throughout my study. Last but not least, my sincere thanks goes to my partner, Karol Castillo for all her unremitting encouragement, patience and continuous support during my PhD study.
Table of Contents

ABSTRACT iii

ACKNOWLEDGEMENTS v

TABLE OF CONTENTS ix

LIST OF TABLES xiii

LIST OF FIGURES xiv

LIST OF PAPERS xvii

ABBREVIATIONS xxi

1 INTRODUCTION 1

1.1 Introduction ................................................................. 1

1.2 Background ................................................................. 2

1.3 Intensive Care Unit ......................................................... 4

1.4 ICU Patient Flow in Hospital .............................................. 8

1.4.1 Indicators to Measure the Quality of Care in ICU ................. 10

1.4.2 ICU Readmission Rate ................................................... 11

1.4.3 ICU and Hospital Length of Stay ..................................... 12

1.4.4 ICU Occupancy Rate ..................................................... 15

1.4.5 ICU Mortality Rate ...................................................... 18
1.5 Critical Care Outreach Team ................................................. 20
   1.5.1 The Team Composition and Size ................................... 22
   1.5.2 Communication Strategy and Teamwork ........................... 24
   1.5.3 Hours of Outreach Services ........................................... 26
   1.5.4 Education and Team Skills ............................................. 27

1.6 Aims and Objectives ............................................................. 28

2 Research Design and Methodology ........................................... 31

   2.1 Introduction ......................................................................... 31

   2.2 Research Design ............................................................... 31
      2.2.1 Literature Review and Gap Analysis ............................... 32
      2.2.2 Designing Research Questions ..................................... 34
      2.2.3 Identify Choice of Research Method ............................... 36
      2.2.4 Ethics Approval .......................................................... 38
      2.2.5 Data Collection and Analysis ........................................ 38
      2.2.6 Verification and Validation of Proposed Models .............. 43
      2.2.7 Recommendation ....................................................... 45

3 Literature Review ................................................................. 47

   3.1 Introduction ......................................................................... 47

   3.2 Intensive Care Unit in Operations Management Literature .... 48
      3.2.1 Background ............................................................... 48
      3.2.2 Modelling of Admission and Discharge Policies in ICU .... 48
         3.2.2.1 ICU Performance and Patient Outcomes .................. 48
         3.2.2.2 ICU Occupancy .................................................. 52
      3.2.3 Modelling of Uncertainties in ICU ................................. 53
         3.2.3.1 Classifying Patients ............................................. 53
         3.2.3.2 Arrival Process and Length of Stay ......................... 54
   3.3 Nurse Staffing and Rostering .............................................. 55
# Table of Contents

3.3.1 *Background* ........................................................................... 55
3.3.2 *Nurse Rostering Problem* ...................................................... 55
3.3.3 *Nurse Scheduling Problem* .................................................... 57

3.4 *Rapid Response Systems* .......................................................... 59
3.4.1 *Introduction* ......................................................................... 59
3.4.2 *Afferent Limb* ...................................................................... 60
3.4.3 *Efferent Limb* ...................................................................... 60
3.4.4 *Administrative and Quality Improvements* ......................... 62

3.5 *Queueing Theory* ................................................................. 62
3.5.1 *Origins, Standard Results and Applications* ....................... 62
3.5.2 *Intractability and Practical Applications* ............................... 67
   3.5.2.1 *Heavy and Light Traffic Approximations* ...................... 67
   3.5.2.2 *Large Deviations Theory* ............................................. 68
   3.5.2.3 *Simulation* ................................................................. 69

3.6 *Regression Analysis* .............................................................. 76
3.6.1 *Linear Regression* ............................................................... 76
3.6.2 *Generalised Linear Model* .................................................... 79

4 *Paper I* ..................................................................................... 85

   4.1 *Introduction* ........................................................................ 86
   4.2 *Literature Review* ................................................................. 89

4.3 *The PART-ICU Interaction Process and Data Description* .... 92
   4.3.1 *PART Process* ................................................................. 92
   4.3.2 *Data Description* ............................................................. 93

4.4 *Measuring the Value of ICU and PART Visits to Patient Outcomes* 99
   4.4.1 *Evaluation of the Marginal Value of a Nurse* .................... 99
   4.4.2 *Econometric Model for Patient Outcomes* ....................... 102

4.5 *Estimation Results* ............................................................... 104
# Table of Contents

## 4.5.1 “What-if” Analysis of Possible Nurse Re-allocation Decisions

4.6 Conclusion and Future Research

## 5 Paper II

### 5.1 Introduction

### 5.2 Literature Review

### 5.3 Model

#### 5.3.1 Process Description

#### 5.3.2 Data Description

#### 5.3.3 Ward-ICU Queueing Model

#### 5.3.4 The PART Queueing Model

#### 5.3.5 Simulation Model

##### 5.3.5.1 Building Simulation Model in Arena

##### 5.3.5.2 Verification and Validation

##### 5.3.5.3 Nurse Staffing Optimisation Problem

### 5.4 Results and Discussion

### 5.5 Conclusion

## 6 Discussion and Conclusion

### 6.1 Discussion

### 6.2 Summary of Research Results

### 6.3 Contribution to Research

### 6.4 Contribution to Practice

## 7 Future Research Direction

## References
List of Tables

2.1 Data Description .................................................................................. 42

3.1 The Advantages and Disadvantages of two Effector Limb Response Team Models 62

4.1 Summary Statistics of Patients With and Without PART Visits ................. 96
4.2 Number of Patients in each PART Episode Outcome ............................... 98
4.3 Description of the Patient Characteristics .............................................. 102
4.4 Description of Operational Variables .................................................... 103
4.5 The Impact of Occupancy Level of ICU Nurses during the Admission Time on Patient Hospital Length of Stay (n = 4334) ........................................... 106
4.6 Number of Patients in Different Statuses of ICU Occupancy .................. 107
4.7 Number of Patients in Different Statuses of PART Occupancy ................. 107
4.8 The Impact of Occupancy Level of PART Nurses on Patient Hospital Length of Stay (n = 710) ................................................................. 108
4.9 Net Changes by Moving Nurses between PART and ICU ...................... 110

5.1 Model Parameters .................................................................................. 130
5.2 Interarrival Time Distribution ................................................................ 133
5.3 Length of Stay Distribution .................................................................... 134
5.4 Validation of Simulation Model by Comparing the Simulation and Real Outcomes ................................................................. 136
5.5 Nurse Allocation Scenarios between PART and ICU ............................ 138
5.6 Number of ICU Deaths in Different Nurse Allocation Policies between PART and ICU ................................................................. 140
## List of Figures

1.1  Intensive Care Unit *(Source: www.aviva.co.uk)* 

1.2  Adult Hospital Stays and Aggregate Total Hospital Charges by Intensive Care Unit (ICU) in the USA 

1.3  The Comparison of nurse-to-patient Ratio in Ward, ICU and SDU 

1.4  The Comparison of nurse-to-patient Ratio in Ward, ICU and SDU 

1.5  Admissions to Adult ICUs by Hospital Classification 

1.6  Source of Admission by Hospital Classification 

1.7  ICU Patient Flow in Hospital: Operating theatre (OT), Post-Anesthesia Care Unit (PACU), Intermediate care (IMC) and Emergency Department (ED) 

1.8  The Average Hospital Length of Stay (2000-2015) 

1.9  After Hours Discharge by Hospital Classification 

1.10  After Hours Discharge by Tertiary Sites 2015/16 

1.11  RRT Specifications by Hospital Type and Size 

1.12  The National Vital Signs Chart 

1.13  The Escalation Pathway 

2.1  Research Design and Methodology Steps 

2.2  Correlation of APACHE II score with Hospital Mortality 

3.1  Rapid Response System Structure 

3.2  Visensia Early Warning System *(Source: www.obsmedical.com)* 

3.3  Application of Queueing Theory in Health care
List of Figures

3.4 A Queueing System ............................................................. 64
3.5 The Position of Discrete-event Simulation and Monte Carlo Simulation .................. 71

4.1 PUP Scoring System (Source: PUP scoring system used in Counties Manukau Health in New Zealand) ........................................... 91
4.2 The Flowchart of PART Interactions with Wards and ICU/HDU ........................................... 93
4.3 The PART Patient Data Diagram ................................................ 94
4.4 Histograms of Different Data Dimensions ........................................ 95
4.5 ICU Nurse Occupancy ................................................................ 95
4.6 Histogram of the PART Occupancy ................................................ 97
4.7 Density of ICU Length of Stay of Patients With and Without PART Visit ..................... 98
4.8 Cumulative Distribution Function of Admissions and Discharges ......................... 100
4.9 ICU Admission Rate under Low and High Occupancy Level of Nurses .................. 101
4.10 Hospital Length of Stay ............................................................. 104

5.1 PART-ICU Process Chart .......................................................... 122
5.2 PART Arrival Distribution .......................................................... 123
5.3 PART Length of Stay ................................................................. 124
5.4 ICU Arrival Distribution ............................................................. 125
5.5 ICU Length of Stay ................................................................. 126
5.6 Ward-ICU Queueing Model ........................................................ 128
5.7 PART Queueing Model .............................................................. 129
5.8 Interarrival Time Distribution for ICU with PART ..................................... 134
5.9 Length of Stay Distribution for ICU with PART ..................................... 135
5.10 Work In Process in Ward Queue .................................................. 136
5.11 Number of Deaths after ICU Discharge .......................................... 138
5.12 ICU Monthly Admission from Different Departments ................................. 139
5.13 Arena Simulation Model ........................................................... 142
List of Papers

This thesis encompasses two original papers, as well as the introduction, research design and methodology, literature review and discussion and conclusion chapters. The papers are listed below with a brief summary of the applied methodology, findings and the author’s contribution.

Paper I


It is vital for a hospital management to efficiently manage limited ICU resources and particularly allocate them to those acutely ill patients who would benefit most from receiving them. In addition to the classical solution of implementing admission and discharge policies, most hospitals have also recently implemented a Patient-At-Risk-Team (PART). The primary goal of PART is to avert ICU admissions by providing critical care to patients in need in a ward. However, PART has to be staffed with ICU-level nurses, but in practice this is often possible by reallocating an ICU nurse to PART. So, the managerial concern is how to split critical care nurses between ICU and PART optimally. On the academic side, this research contributes to the staffing problem in the acute care setting. Our analytical approach also helps managers to derive more meaningful insights from the available data. We use econometric models to estimate the marginal values of a nurse at ICU and at PART,
and use those values along with occupancy data to estimate the effect of moving one nurse from ICU to PART and vice versa. We found that when the ICU nurses are busy the patient hospital length of stay increases by 3% which leads to an increase in the overall patient hospital length of stay by nearly 125 patient-days. We also figured out that the high utilisation rate of the PART nurses expands the patient hospital length of stay by almost 70 patient-days. We evaluated different scenarios of nurse staffing between ICU and PART and found that it is more beneficial to the hospital to move a nurse from PART to ICU. Besides, we figured out that the hospital manager could save approximately $300,000 per annum by allocating a new nurse to PART per shift.

As the first author, Ali Haji Vahabzadeh has taken the lead to develop the models, collect the data and write the paper, with contributions made by the co-authors: Dr Valery Pavlov, who is the main supervisor for the doctoral research, and who clarified the models, editorial matter and explanatory writings and Dr Alex Kazemi, who is the ICU consultant at Middlemore Hospital and provided us with the required data and descriptions of the medical terms and processes.

**Paper II**


Some critically ill patients in a hospital ward may receive sub-optimal care because their deterioration is not identified or not acted upon timely by the ward nurses. At the same time, a low-risk patient might be wrongly identified by a nurse in the ward as an acutely ill patient who needs to be admitted to an ICU. To detect the early signs of critical health deterioration among patients in the ward, and also avert preventable ICU admissions, hospitals implement a nurse-led PART. However, today’s global nursing shortage causes a serious and complex challenge for hospital managers on determining nurse staffing levels
in hospital’s medical units. In this paper, we use a queueing model to examine the impact of the nurse staffing level between PART and ICU on the ICU patient mortality rate. Using operational flow of patients of a large hospital in New Zealand, we build a discrete-event simulation model. Our results suggest that by reallocating even one nurse from ICU to PART, the ICU mortality rate can decrease by as much as 35%. The results also illustrate why, despite such a dramatic potential benefit for patients, a decision to re-allocate even one nurse from ICU to PART can be difficult for managers. The reason is that the utilisation of PART nurses tends to be much lower, by an order of magnitude, than utilisation of ICU nurses. Thus, our study sheds light both on the significant role of PART in hospitals on improving the patient outcomes and ICU performance and also provides some insight into the type of challenges presented to hospital managers.

As the first author, Ali Haji Vahabzadeh has built the queueing and simulation models with contributions coming from the main supervisor, Dr Valery Pavlov, who has made recommendations on model construction and calibration.
Abbreviations

**AHP**  Analytic Hierarchy Process.

**ALF**  Afferent Limb Failure.

**ANZICS**  Australian and New Zealand Intensive Care Society.

**APACHE**  Acute Physiology and Chronic Health Evaluation.

**CCOT**  Critical Care Outreach Team.

**CCU**  Coronary Care Unit.

**DES**  discrete-event simulation.

**EC**  Emergency Care.

**ED**  Emergency Department.

**EHR**  Electronic Health Record.

**ESPI2**  Early Detection of Impending Physiologic Deterioration, version 2.

**EWS**  Early Warning Score.

**FAWAS**  Feature Analysis and Weighted Average Sum.

**FCFS**  First-Come-First-Served.

**FICM**  Faculty of Intensive Care Medicine.
About this resource

This resource is designed to support the provision of optimal nursing care in the ICU. It provides a comprehensive guide to nurse staffing in the ICU, focusing on the development of effective nurse staffing models. The resource includes an introduction to the nurse staffing problem, an overview of the methods used to develop nurse staffing models, and a detailed discussion of the nurse staffing model developed for the ICU at Royal Darwin Hospital.

The resource also includes an extensive list of abbreviations used throughout the document, including FTE, GA, GLM, HDU, ICU, IMC, ITU, JFICM, LOS, MDP, MET, MLE, NHS, NICU, NP, NRP, NSP, NZEWS, and NZNO.

Conclusion

The development of effective nurse staffing models is crucial for ensuring optimal care in the ICU. This resource provides a comprehensive guide to nurse staffing in the ICU, including an overview of the methods used to develop nurse staffing models and a detailed discussion of the nurse staffing model developed for the ICU at Royal Darwin Hospital. The inclusion of an extensive list of abbreviations used throughout the document is also a valuable resource for healthcare professionals.

Acknowledgements

The author would like to thank the staff at Royal Darwin Hospital for their support and assistance in developing the nurse staffing model. The author would also like to thank the NZNO for their support and assistance in the development of this resource.
Abbreviations

OECD  Organisation for Economic Co-operation and Development.

OLS  Ordinary Least Square.

OM  Operations Management.

OR/MS  Operations Research/Management Science.

OT  Operating Theatre.

PACU  Post-Anaesthesia Care Unit.

PART  Patient-AT-Risk-Team.

PICU  Paediatric Intensive Care Unit.

PUP  Physiologically Unstable Patient.

RCU  Respiratory Care Unit.

RN  Registered Nurse.

RRS  Rapid Response Systems.

RRT  Rapid Response Team.

RRTM  Rapid Response Team Member.

RSS  Residual Sum Square.

RT  Respiratory Therapist.

SA  Simulated Annealing.

SCCM  Society of Critical Care Medicine.

SD  System Dynamic.

SDU  Step Down Unit.
Abbreviations

**SET** Surgical Emergency Team.

**TS** Tabu Search.

**WIP** Work In Process.
Chapter 1

Introduction

1.1 Introduction

This thesis studies the nurse staffing policies in a critical care setting in hospitals. We specifically examine the impact of different critical care nurse allocations between an **Intensive Care Unit (ICU)** and a **Patient-AT-Risk-Team (PART)** on both ICU and ward patient outcomes in one of the largest hospitals in New Zealand. In summary, we make the following contributions: Firstly, we develop econometric models to estimate the impact of occupancy levels of critical care nurses in both PART and ICU on patient hospital length of stay. Secondly, we propose a queueing model of PART and ICU to compare different scenarios of nurse allocation between PART and ICU aiming at minimising ICU mortality rate. Lastly, we validate our proposed model using a 12-month history (1 July 2015 to 30 June 2016) of 8,576 visits of the PART to 2,662 patients across all wards and post-ICU admission outcomes of patients admitted via referral from PART.

The remainder of this introductory chapter is organised as follows: Section 1.3 describes the background of the problem and our motivations to perform this study. Section 1.3 introduces the ICU, and its role in a patient journey in a hospital. We highlight the ICU patient flow in hospitals and the common Key Performance Indicators (KPIs) in the ICU in Section 1.4. We further discuss different outreach services in a critical care context in...
hospitals and analyse the associated factors impacting their performance in Section 1.5. Finally, having introduced the problem, we present the aims and objectives of our thesis in Section 1.6.

1.2 Background

An ICU is a medical unit of a hospital that provides the most expensive and invasive forms of care to acutely ill patients in life-threatening conditions. It is equipped with exclusive monitoring equipment and specially-trained staff that provide constant support to the ICU patients. Not surprisingly, it is also very costly. Approximately 15% to 20% of the total hospital expenses [Berenson & Assessment, 1984], and 4.1% of the US health expenditures [Halpern & Pastores, 2010] are spent in the ICU while ICU’s beds contain less than 10% of the total hospital beds [Edbrooke, Hibbert, Ridley, Long, & Dickie, 1999].

Due to the acuity and complexity of ICU patients, ICUs provide round the clock nurse supervisory to each patient so that typical nurse-to-patient ratios are much higher than those in other units (typically, 1:1 at the ICU, 1:2 in a step-down unit and 1:6 in a ward). However, maintaining the high number of highly qualified nurses in the ICU is challenging for hospital managers as it is mostly influenced by the global nursing shortage. For example, in the US, a need for more than 3.4 million nurses is projected by the American Nurse Association by 2022. In 2012, the New Zealand Nurses Organisation (NZNO) also reported shortages of over 120 Registered Nurse (RN) positions at Auckland District Health Board (DHB) alone. Likewise, the Faculty of Intensive Care Medicine (FICM) in the UK reports that 96% of acute hospitals operated without an adequate level of nursing staff during day shifts in October 2016 [Silva, 2017].

Considering the cost and scarcity of ICU nurses, it is imperative to utilise them in the most effective way. However, the results of medical studies reveal that about 40% of ICU admissions are potentially preventable [McQuillan et al., 1998]. Specifically, the discussions in the medical literature demonstrate that many patients in the ward are acutely ill and
require immediate critical care services, however, their physiological deteriorations are either not recognised by the ward staff on time or not acted upon correctly resulting in disastrous and irreparable consequences such as unplanned ICU admissions, prolonged ICU Length of Stay (LOS) or even patient death.

To timely respond to the urgent needs of critically ill patients in wards, on the one hand, and to prevent unplanned ICU (re)admissions, on the other hand, a Critical Care Outreach Team (CCOT) in hospitals was recommended in the UK in 2000 (DoH, 2000). The primary objectives of CCOT were outlined as: to avert unplanned ICU admissions by appropriately identifying the deteriorating patients on wards, to facilitate ICU discharge by providing follow-up supports on wards and finally to share critical care skills with junior staff on wards. Similar functionality but different sizes and compositions (e.g., nurse-led or medically-led) of the outreach services have also been implemented in hospitals throughout the world: a Medical Emergency Team (MET) in Australia (Lee, Bishop, Hillman, & Daffurn, 1995), a Patient-At-Risk Team (PART) in New Zealand (Franklin & Mathew, 1994) and a Rapid Response Team (RRT) in the USA (K. Thomas, VanOyen Force, Rasmussen, Dodd, & Whildin, 2007).

In spite of massive efforts in constructing the outreach services in hospitals, the results of medical studies on the efficacy of the outreach services on patient outcomes and ICU performance are still equivocal (we discuss the evidence in detail in the Literature Review section). In fact, it has not yet been well-established how and when constructing a costly outreach team in hospitals while managers are struggling with supplying the scarce critical care nurses to ICU, is beneficial to both patients and the hospital. Therefore, to deliver high-quality and cost-effective care to both severely ill patients in the ward and ICU patients, it is vital for hospital managers to ration the limited number of critical care nurses between PART and ICU in the most efficient way. We will pursue our discussions on optimal nursing between PART and ICU in the following chapters.
1.3 Intensive Care Unit

In 1997, the Joint Faculty of Intensive Care Medicine (JFICM) defined an ICU as “a specially staffed and equipped, separate and self-contained section of a hospital for the management of patients with life-threatening or potentially life-threatening conditions. Such conditions should be compatible with recovery and have the potential for an acceptable future quality of life. An ICU provides special expertise and facilities for the support of vital functions, and utilises the skills of medical nursing and other staff experienced in the management of these problems” (Figure 1.1).

![Intensive Care Unit Diagram](Source: www.aviva.co.uk)

The ICU is also known as an Intensive Therapy Unit (ITU) in some hospitals throughout the world. Acutely ill patients are mainly admitted to the ICU due to a failure of one
or more of their body’s organ systems such as the heart, lungs or even the brain. The reasons why patients are admitted to the ICU vary extensively, ranging from a heart attack and stroke to surgical complications and a car accident. Most ICU patients are connected to different monitoring equipment and life-supporting machines such as a mechanical ventilator as they may have some difficulties with their breathing. The structure of ICUs also differs based on the medical speciality or patient type (e.g., infant or adult). For example, a Neonatal Intensive Care Unit (NICU) provides services to neonatal patients who have not yet been discharged from the hospital after birth, or a Paediatric Intensive Care Unit (PICU) and Coronary Care Unit (CCU) that are mainly customised for acutely ill paediatric and cardiac arrest patients, respectively.

Figure 1.2 Adult Hospital Stays and Aggregate Total Hospital Charges by Intensive Care Unit (ICU) in the USA (Source: Barrett, Smith, Elixhauser, Honigman, & Pines, 2014)

Due to the use of highly trained staff (doctors and nurses) and advanced equipment, the access to ICU services is solely provided to acutely ill patients who would benefit most
from such services (e.g., post-surgical or multiple organ failure patients). Indeed, due to limited and expensive resources, ICU is reluctant to admit patients with higher risk of severe complications such as haematological (Moors & Benoit, 2014) or very old patients (Nguyen, Angus, Boumendil, & Guidet, 2011).

Concerning the ICU cost, it has been widely reported that ICU is one of the most expensive medical units in the hospital. The ICU usually costs four (Rapoport, Teres, Zhao & Lemeshow, 2003) to six (Griffiths, Price-Lloyd, Smithies, & Williams, 2005) times that of a general ward. An analysis of patient’s hospital stays in 29 states of the US in 2011 illustrates that hospital stays with ICU admissions were two and a half times more costly than other hospital stays (Figure 1.2). The total cost per patient per ICU day is roughly €855 (inflated to 2008) in Germany (Moerer et al., 2007), €3221 (inflated to 2008) in the US (L. M. Cooper & Linde-Zwirble, 2004) and $5000 in New Zealand (https://www.pharmac.govt.nz/). The type of ICU (e.g., surgical or medical), patient mix (Angus et al., 2006; P. Kumar, Jithesh, & Gupta, 2015) and interventions such as mechanical ventilation (Dasta, McLaughlin, Mody & Piech, 2005) may impact the daily costs in the ICU. For example, almost 33% of ICU patients require mechanical intervention which incurs a high share of the total cost of ICU treatment (Esteban et al., 2002; Hébert et al., 2001).

In cases where patients require more intensive observation or treatment than in a general ward but less than the care provided in the ICU, another intermediate unit called a High Dependency Unit (HDU), or a Step Down Unit (SDU) accommodates the candidate patients (e.g., patients with single-organ failure). Gotsman and Schrire (1968) initially proposed the concept of HDU for cardiac patients who no longer require intensive care services but are still not healthy enough to get discharged to a general ward. According to Prin and Wunsch (2014), HDU patients can be categorised into three groups: 1. patients admitted from ICU (usually organ support) and no longer require full ICU services, 2. patients admitted from an Emergency Department (ED) or general ward who need more care and observation, and 3. postoperative patients transferred from an Operating Theatre (OT) or a recovery room.

It is evident that because of the interdependency of services provided to patients
in both ICU and HDU, the two units should also be physically adjacent to each other. Some hospitals establish a combined ICU and HDU where a certain number of beds are allocated to HDU patients with a slightly lower level of care than the ICU beds, while some other hospitals set up a standalone HDU but near the ICU. Figure 1.3 represents different locational relationships between the general ward, HDU and ICU. Although the model of co-located units requires flexible nursing (Chaboyer, James, & Kendall, 2005; Heller & Murch, 2008), which may be of interest to some hospital managers, low acute patients may suffer from sleep deprivation due to nursing interventions and disruptive noise (Darbyshire & Young, 2013).

ICU and HDU also differ in the nurse-to-patient ratios. Despite the ratios varying in different ICUs, the nurse-to-patient ratios are usually 1:1 or 1:2 in ICUs and 1:2 or 1:3 in HDUs. Notwithstanding, these ratios could be largely influenced by factors such as budgetary considerations (Clarke & Donaldson, 2008), shortage of critical care nurses (Kwiecień, Wujtewicz, & Mędrzycka-Dąbrowska, 2012; Spence et al., 2006) and patient acuity and complexity of care (Penoyer, 2010). Prin and Wunsch (2014) also illustrate the impact of physical location of the HDU with regards to the ward and the ICU on the nurse staffing policies. As presented in Figure 1.4 for instance, an SDU combined with an ICU...
may lead to a nurse-to-patient ratio of 2:5, while an SDU integrated with a general ward may even result in a nurse-to-patient ratio of 2:11.

1.4 ICU Patient Flow in Hospital

In the previous section, we only discussed patient flows between ICU and HDU. Nonetheless, the ICU also provides acute services to upstream (e.g., OT and ED) and down-stream (e.g., wards, Post-Anaesthesia Care Unit (PACU) and Intermediate Care Unit (IMC)) departments. In the US, for example, more than 5.7 million patients are admitted to ICUs per annum (Barrett et al. 2014), of which usually 28% to 43% are admitted to the ICU from OT, 36% to 42.5% from ED, 11% to 18% from ward and 5% to 11% from other ICUs or hospitals (Groeger et al. 1993; Knaus, Wagner, Zimmerman, & Draper 1993; Sirio et al. 1999; Zimmerman et al. 1998). In New Zealand, the majority of ICU patients are also transferred from ED (42%), while OT and ward patients account for 18% of ICU admissions.
In a broader view, in 2014, the Australian and New Zealand Intensive Care Society (ANZICS) published the source of ICU admissions according to different hospital’s classes based on hospital location, available services and case mix in both Australia and New Zealand (Figure 1.5). As Figure 1.6 shows, in 2015/16, private ICUs admitted a higher proportion of patients from OT or those who required recovery following an elective surgery. In Metropolitan and rural hospitals, ED patients account for the largest percentage of ICU admissions. And, finally, among the public hospitals (tertiary, metropolitan and rural), tertiary ICUs served the highest number of acutely ill patients transferred from OT and recovery.

Besides, as shown in Figure 1.7, some patients after surgery, with or without going through PACU, might be directly sent to the ICU as they need special requirements for monitoring of their blood pressure or respiratory rate. Note that, a high level of collaboration and engagement is particularly required to be designed between ICU and OT. In fact, owing to the capacity constraints (bed and nurses) in the ICU, some post-surgical patients might be denied ICU admission, resulting in reduction in patient throughput (McConnell et al., 2005). Conversely, the ICU will not be able to provide the required services if extra surgeries are scheduled without taking the scarce capacity of the ICU into account (Fügener et al., 2016).
Once ICU patients reach some level of stability, they will be either discharged to the ward directly or admitted to IMC (e.g., HDU) if it is felt that they still need more observation. The IMC or ward patients could be further sent back to the ICU if it was realised that their health condition had deteriorated. The majority of ICU patients are finally discharged to the wards or HDU rather than home (there are also some cases of ICU patients transferring to other ICUs or hospitals. This is mainly because the patient needs specialist care but it is not available in the current ICU).

### 1.4.1 Indicators to Measure the Quality of Care in ICU

A large number of clinical studies discuss the importance of quality of services in the ICU (Angus & Black, 2004; Bion & Heffner, 2004; Cook, Montori, McMullin, Finfer, & Rocker, 2004) and propose some quality indicators that measure and monitor the structure, process and outcome of care (Donabedian, 1973) as follows:

- **Structure indicators:** According to Curtis et al. (2006), structure indicators mainly focus on the way we identify ICU care, size, resources and their integration with
other departments. The ICU nurse staffing or availability of ICU consultants are two examples of structure indicators.

- Process indicators: Process indicators are the second element of the quality of care that quantifies what the healthcare provider did for the patient and how well the service was delivered (Mainz, 2003). The patient’s ICU length of stay and ICU occupancy rate (Berenholtz, Dorman, Ngo, & Pronovost, 2002) are the two common process indicators that are being measured by the majority of both clinical and Operations Management (OM) studies.

- Outcome indicators: Those indicators, which measure the effects of services on the health and wellbeing of patients, are categorised as outcome indicators (e.g., ICU mortality and readmission rates).

We will look further into those indicators that have been used in most OM research.

### 1.4.2 ICU Readmission Rate

ICU readmission rate is defined as the percentage of patients readmitted to the ICU within 72 hours of ICU discharge within the same hospitalisation (Chrusch, Olafson, McMillan, Roberts, & Gray, 2009) (Note though, that some studies such as Angus (1998) defined the ICU readmission within 48 hours of ICU discharge). About 4% to 10% of ICU patients are normally readmitted to the ICU between 24 and 72 hours after the first discharge from the ICU (Frost et al., 2009, 2010; Hosein et al., 2014). The ICU readmitted patients tend to have a longer ICU length of stay (ICU LOS) compared with their first admission (4.9 versus 3.4 days) (Kramer, Higgins, & Zimmerman, 2013; J. A. Russell, 2012; Santos et al., 2014). Their mortality risk is almost 21% higher, as well (Chrusch et al., 2009; Kramer et al., 2013; J. A. Russell, 2012).

The findings reveal that the main predictors of the ICU readmission rate are age, gender, severity of illness, admission type, comorbidities and length of stay (Chrusch et al., 2009; Frost et al., 2009, 2010; Giakoumidakis et al., 2014; Kramer et al., 2013; J. A. Russell, 2012).
Santos et al., 2014), ICU first discharge during weekends and nights (S. E. S. Brown, Ratcliffe, Kahn, & Halpern, 2012; Chrusch et al., 2009; J. A. Russell, 2012), and mechanical ventilation in the first ICU admission (Frost et al., 2009; Santos et al., 2014; Wagner et al., 2013).

The unplanned ICU readmission is also significantly associated with higher cost and resource utilisation (Desautels et al., 2017; Lai et al., 2012). As the considerable proportion of ICU expenses are accumulated on the first day of admission, the cost of ICU readmission is supposed to be a second ‘first day’ of ICU admission (S. Russell, 1999). It is widely argued that patients readmitted within 48 hours of ICU discharge may have been prematurely discharged from the ICU (Rosenberg & Watts, 2000). This is mainly because an early discharge of a stable ICU patient not only could reduce the costly prolonged ICU stay but also could free up some invaluable capacities of ICU for other critically ill patients (C.-L. Liu et al., 2017; Rosa et al., 2015). Nonetheless, the early discharge from ICU is too risky especially for acute patients as it can result in either further ICU readmission or even patient death (Rosenberg & Watts, 2000). Therefore, designing an optimal ICU discharge policy could assist decision makers to reduce the ICU readmission rate and improve the ICU capacity simultaneously.

1.4.3 ICU and Hospital Length of Stay

Length of stay (LOS) is a standard quality indicator used to compare the performance and quality of services among hospitals (J. W. Thomas, Guire, & Horvat, 1997). It is highly correlated with hospital costs (Cots, Elvira, Castells, & Sáez, 2003; Polverejan, Gardiner, Bradley, Holmes-Rovner, & Rovner, 2003) which is why hospitals attempt to shorten the unnecessary patient LOS (Siskou, Kaitelidou, Economou, Kostagiolas, & Liaropoulos, 2009). This fact is evident in the Organisation for Economic Co-operation and Development (OECD)’ health data (Figure 1.8) that illustrates how the average hospital LOS (HOSLOS) decreased in different countries between 2000 and 2015. It is worth mentioning that the HOSLOS depends on various clinical and demographic factors. For instance, acute myocar-
dial infarction patients may stay in a hospital between 4.17 and 41 days while the HOSLOS for C-section patients could range from 2.7 to 16 days (Hasan, Orav, & Hicks, 2010).

For the acutely ill patients, in particular, ICU LOS along with patient illness severity and mortality are the most common variables measured to assess the performance and efficacy of the quality of services in the ICU (Marik & Hedman, 2000). In general, the main goal of the ICU is to lessen the patient LOS without jeopardising their health, leading to a reduction in healthcare cost and unnecessary use of limited resources (Gruenberg et al., 2006). There are many ways to calculate the ICU LOS such as the number of calendar days (LOS-calendar), midnight bed-occupancy days (LOS-midnight) or exact LOS calculated in hours divided by 24 (LOS-exact) (Marik & Hedman, 2000), but ICU LOS is typically calculated based on the time difference between ICU admission date and time and ICU discharge date and time (Verburg, de Keizer, de Jonge, & Peek, 2014).

![Figure 1.8 The Average Hospital Length of Stay (2000-2015) (Source: OECD, 2017)
Studies usually report the average LOS (Becker et al., 1995; Goins, Reynolds, Nyanjom, & Dunham, 1991; MacKenzie, Morris, & Edelstein, 1989), but the frequency distribution of LOS is often right-skewed (Weissman, 1997) as the right tail represents patients with prolonged LOS. It is also discussed that arithmetic means are largely affected by outliers and are not suitable for explaining the “typical” length of stay (Maxfield, Schweitzer, & Gouvier, 1988; Ratcliff, 1993) as opposed to mode and median. Nonetheless, in very long-tailed distributions, the median may also not accurately describe the central tendency (Ratcliff, 1993). In such cases, the harmonic and geometric tend to be better measurements than the median (Weissman, 1997).

While the average ICU LOS is 3.3 days (Hunter, Johnson, & Coustasse, 2014), there is no standard definition for a prolonged ICU LOS (Weissman, 1997) as it alters by the type of hospital, ICU, and also patient sickness (Laupland, Kirkpatrick, Kortbeek, & Zuege, 2006; C. M. Martin, Hill, Burns, & Chen, 2005; Trottier, McKenney, Beninati, Manning, & Schulman, 2007). For example, Weissman (1997) defined a prolonged ICU stay as more than 21 days at teaching hospitals and more than 10 days at community hospitals. Or, while Mahesh et al. (2012) defined the ICU stay of more than three days as prolonged ICU LOS, Huang, Huang, Tsauo, and Ko (2010) determine it as when the ICU LOS exceeds 16 days because rates of ICU mortality, hospital mortality and mortality one year after ICU discharge remain constant after ICU LOS was more than 16 days.

It has been particularly projected that 2% to 11% of acutely ill patients require a prolonged stay in ICU (Delle Karth, Meyer, Bauer, Nikfardjam, & Heinz, 2006; Gersbach et al., 2006; Hein et al., 2006; Laupland et al., 2006; C. M. Martin et al., 2005) constituting 25% to 45% of total ICU days (Arabi, Venkatesh, Haddad, Al Shimemeri, & Al Malik, 2002; C. M. Martin et al., 2005; Wong, Gomez, McGuire, & Kavanagh, 1999) and using a considerable amount of ICU resources (Stricker, Rothen, & Takala, 2003; Weissman, 2000; Wong et al., 1999; Zilberberg, Luippold, Sulsky, & Shorr, 2008). Patients who stay more than 30 days in ICU utilise about 16% of total ICU bed days, even though they account for almost 1.5% of ICU admissions (Hughes, MacKirdy, Norrie, & Grant, 2001; Stricker et al.)
also found that while only 11% of patients stay more than seven days in ICU, they employ 50% of ICU resources.

The results of clinical works also reveal that the prolonged ICU LOS is associated with higher mortality and morbidity rates (Rajakaruna, Rogers, Angelini, & Ascione, 2005; Rimachi, Vincent, & Brimioulle, 2007). Explicitly, the mortality of patients stay in the ICU for two weeks or more was predicted to be approximately 50% (T. A. Ryan et al., 1997; Wong et al., 1999). In addition, 70% of patients with ICU LOS of more than two weeks had less than 50% functional recovery (Fakhry, Kercher, & Rutledge, 1996). Thus, it is of paramount interest to identify patients with higher risk of prolonged ICU stay as that may improve both resource utilisation and survival of ICU patients.

1.4.4 ICU Occupancy Rate

As we argued in the previous section, ICU LOS accounts for a considerable segment of the total hospital cost. One of the operational variables that impacts the ICU LOS is the occupancy level of the ICU (Kc & Terwiesch, 2011; S.-H. Kim, Chan, Olivares, & Escobar, 2014). In general, workload can be defined as “the relative capacity to respond” (Lysaght, 1989). “Workload is a construct that is used to describe the extent to which an operator has engaged the cognitive and physical resources required for a task performance” (Backs, Ryan, & Wilson, 1994). “Workload is a multidimensional and complex construct, that is affected by external task demands, environmental, organisational and psychological factors, and perceptive and cognitive abilities” (Weinger, Reddy, & Slagle, 2004). Based on these definitions, workload encompasses three main elements: (1) an operator, using their resources to respond to (2) external physical or cognitive demands to (3) carrying out a specific task (Hoonakker et al., 2011).

In particular, the discussion on ICU occupancy can be categorised into two streams: ICU nurse occupancy and ICU bed occupancy. Studies that take ICU nurse utilisation into account mainly analyse the impact of nurse staffing on patient outcomes (Aiken, Clarke, Sloane, Lake, & Cheney, 2008; Tourangeau, Cranley, & Jeffs, 2006; Unruh, 2008).
such as mortality (Aiken et al., 2008; Aiken, Clarke, Sloane, Sochalski, & Silber, 2002; Blegen, Goode, & Reed, 1998; Needleman, Buerhaus, Mattke, Stewart, & Zelevinsky, 2002), adverse events, complications, failure to rescue (Aiken et al., 2002; Cho, Ketefian, Barkauskas, & Smith, 2003; Needleman, Buerhaus, Stewart, Zelevinsky, & Mattke, 2006), quality of care (Sochalski, 2004), costs (Aiken et al., 2008; Cho et al., 2003; Needleman et al., 2006) and length of stay (Cho et al., 2003; Needleman et al., 2006). These studies apply different ways of measuring nurse staffing. For example, Cho, Hwang, and Kim (2008) use the ratio of average daily census to the total number of Full-time equivalent (FTE) RN. While Kiekkas et al. (2008) defined nursing as patient “exposure” to nursing workload by calculating a ratio of patient demands (Therapeutic Intervention Scoring System-28 sum) to the daily number of nurses over three levels.

Other works, however, study the utilisation rates of ICU beds and apply a different formula to compute this ratio. For example, Parker, Wyatt, and Ridley (1998) obtain the ICU occupancy as the sum of the proportions of the durations of admission in each month relative to the number of available beds. Halpern, Pastores, Thaler, and Greenstein (2006) employ a different approach and calculate the ICU occupancy as the number of bed days used in the ICU divided by the number of licensed bed days (licensed beds are the maximum number of beds approved by the licensing agency, and are not necessarily existent beds (Phillip, Mullner, & Andes, 1984)).

It is noted that the actual ICU occupancy might be underestimated as beds can be out of service due to cleaning, maintenance, patient isolation and even changes in resource scheduling (L. V. Green, 2002; McManus, Long, Cooper, & Litvak, 2004; Rocker, Cook, Martin, & Singer, 2003). In some cases, patients cannot be even included in the calculations. For instance, a patient who was admitted to and discharged from the ICU on the same day might be dropped from the calculations if the ICU LOS is calculated only based on the admission and discharge dates (Tierney & Conroy, 2014). Similarly, the census methods used for calculating the ICU utilisation rate may neglect patients with short ICU LOS (Ridley & Rowan, 1997). Especially, midnight census probably generates the minimum ICU
utilisation rate of the day as patients who are admitted in the morning and discharged at night are excluded from the calculation (L. V. Green 2002).

On the contrary, some methods can result in an over-estimation of utilisation. For example, when a patient ICU LOS is rounded up to a closest full day such as calendar date or day-to-day (Ridley and Rowan, 1997) or where the midday census is used for computing the ICU occupancy. For instance, a patient who is admitted to ICU at 11:30 am on day one and discharged at 12:30 pm on the next day would be considered in the occupancy calculation on both days even though the ICU LOS is just over one 24-hour period (Ridley & Rowan 1997).

ICUs usually report their occupancy levels as an annual average (Tierney & Conroy 2014). Nonetheless, the daily, monthly and seasonal trend of the occupancy levels of ICUs would be more beneficial to policy makers (Barado et al. 2012; Costa et al. 2003; L. V. Green 2002; Parker et al. 1998) as the details of variations in ICU occupancy are more tractable (L. V. Green 2002), although Iwashyna, Kramer, and Kahn (2009), in studying 108 ICUs in 46 hospitals in the US find no significant changes in the patient mortality with increasing census. The consequences of high utilisation rates of ICU are counted as declined or delayed ICU admissions, cancellation of elective surgery, increased severity of illness on later admission to the ICU, nursing critically ill patients elsewhere in the hospital, transfer of emergency patients, hospital diversions, and premature discharge (Barado et al. 2012; Costa et al. 2003; A. Green & Edmonds 2004).

To avoid such negative outcomes, it is suggested that the optimal occupancy in the ICU is approximately 70% to 75% (L. V. Green 2002; Valentin, Ferdinande, & ESICM Working Group on Quality Improvement 2011). This is mainly because maintaining 100% occupancy in ICUs is not achievable as some time is needed for discharging the current ICU patients as well as resetting the beds for the new admissions (Ridley & Rowan 1997). Besides, it is reported that the high utilisation rates of ICUs (usually above 80%) is associated with hospital mortality and ICU re-admission (Chrusch et al. 2009; Iapichino et al. 2010). It might be argued that this range of occupancy for such a costly unit is low. However, it
should be noted that a great proportion of ICU patients are sourced from ED whose arrival times are uncertain. Keeping the ICU occupancy high could thereby turn away some acutely ill patients requiring ICU services. Besides, some ICU beds should be reserved for planned admissions transferred from OR. Thus, maintaining the ICU occupancy high may not be beneficial to hospitals as it could harm reputations as well as referrals.

1.4.5 ICU Mortality Rate

ICU mortality rate is defined as the number of patient deaths in the ICU divided by the number of ICU discharged patients. The average ICU mortality rate varies in different countries. In the US, for example, it ranges from 8% to 19%, or approximately 500,000 deaths annually (Mukhopadhyay et al., 2014). Compared to the US, the ICU mortality rates are lower in Australia (5.04%) and New Zealand (6.86%) (Adult Patient Database Activity Report 2015/2016, 2014). According to the Fact sheet: critical care statistics (2017) released by the Society of Critical Care Medicine (SCCM), multi-organ failure, cardiovascular failure, and sepsis are the primary causes of deaths in ICU, while in Australia and New Zealand, patients with cardiac arrest, sepsis and other ICU infections (not pneumonia) are the highest and lowest diagnostic groups for the ICU fatality, respectively (Adult Patient Database Activity Report 2015/2016, 2014).

The SCCM’s figures show that multi-organ failure has a mortality rate of up to 15% to 28%, new-onset renal failure has a mortality rate of up to 61% (Druml, Lenz, & Laggner, 2015), and severe respiratory failure accounts for 20% to 50% of ICU deaths. Sepsis, the second leading cause of death in ICUs after multi-organ failure and cardiovascular failure, constitutes up to 45% of deaths. Patients with sepsis occupy roughly 50% of ICU resources (Angus et al., 2001; Brun-Buisson et al., 1995; Pittet et al., 1995) incurring an annual cost of $17 billion to the US healthcare system (Angus & van der Poll, 2013; Hall, Williams, DeFrances, & Golosinskiy, 2011). More alarming, the incidence of sepsis has risen over the last two decades, owing to ageing populations, the emergence of drug-resistant pathogens,
and increased use of immunosuppressive drugs (Angus et al., 2001; G. S. Martin, Mannino Eaton, & Moss, 2003).

There are different medical and demographic factors associated with ICU mortality such as age (Chelluri, Pinsky, & Grenvik, 1992; Fuchs et al., 2012; Wood & Ely, 2003), gender (Combes, Luyt, Trouillet, Nieszkowska, & Chastre, 2009; Mahmoud, Eldeirawi, & Wahidi, 2012; Pietropaoli, Glance, Oakes, & Fisher, 2010), illness severity (Y. C. Chen et al., 2001; Girou, Stephan, Novara, Safar, & Fagon, 1998; Lemeshow et al., 1994; G. Li et al., 2016), ICU-acquired infections (Magnason et al., 2008; Ylipalosaari, Ala-Kokko, Laurila, Ohtonen, & Syrjälä, 2006) and mechanical ventilation (Sudarsanam, Jeyaseelan, Thomas, & John, 2005; Vincent et al., 2002).

In addition to all the illness severity related causes, the day and time of admission to and discharge from the ICU might also contribute to ICU fatalities. Several studies reported that after-hours discharge from the ICU to the ward is independently associated with an increased risk of mortality and ICU readmission (Goldfrad & Rowan, 2000; D. A. Harrison, Gao, Welch, & Rowan, 2010; Pilcher, Duke, George, Bailey, & Hart, 2007; Tobin & Santamaría, 2006). As depicted in Figures 1.9 and 1.10 in Australia and New Zealand also a wide variation of after-hours discharge across the 39 Tertiary ICUs from 1.3% to 58.9% (mean of 16.7%) was reported. Finally, the so-called “weekend effect” on ICU patient mortality was
examined. Despite the evidence on increasing the risk of fatality following weekend ICU admission in some research (Barnett, Kaboli, Sirio, & Rosenthal, 2002; Bhonagiri, Pilcher, & Bailey, 2011; Uusaro, Kari, & Ruokonen, 2003), other studies lack sufficient justification of any weekend effects (Arabi, Alshimemeri, & Taher, 2006; Ensminger et al., 2004; Laupland, Shahpore, Kirkpatrick, & Stelfox, 2008) after adjusting for patient illness severity.

### 1.5 Critical Care Outreach Team

ICUs have been experiencing ever-increasing demand. In the US, for example, ICU admissions only from ED dramatically rose by 50% between 2002 and 2009 (Mullins, Goyal, & Fines, 2013). Aging populations and improvements in the delivery of critical care services and organ support techniques are counted as the main reasons for such growth (Crippen, 2013). To meet this demand in reality, the solution of supplying more ICU beds and nurses is confined by the limited budget and shortage of trained staff (Riley & Faleiro, 2001). However, identifying unplanned and potentially preventable ICU admissions could be considered as a practical solution to the better management of ICU demand. This is the fundamental principle of the critical care outreach services to promote the preventive approach in hospitals.

The importance and necessity of implementing outreach services in hospitals were even more realised when the medical results revealed that 54% of ICU patients received suboptimal care before admission to the ICU and about 40% of admissions were potentially preventable (McQuillan et al., 1998). Delays in identification, treatment or referral of deteriorating patients are defined as the main attributes of suboptimal care (Quirke, Coombs, & McEldowney, 2011). A large number of studies have also found preventable deteriorations in patient health conditions prior to ICU admission (Goldhill, 2001; Goldhill, White, & Sumner, 1999; McGloin, Adam, & Singer, 1999; McQuillan et al., 1998).

To address these concerns, a new initiative has widely emerged in hospitals all over the world: in Australia as MET (Lee et al., 1995), in New Zealand as PART (Franklin & Mathew,
Critical Care Outreach Team (1994), in the US as RRT (K. Thomas et al., 2007) and in the UK as CCOT (DoH, 2000). An activation of RRT mainly depends on vital signs disorders (e.g., respiratory or cardiac arrest) detected by the ward nurses as well as their concerns about the health conditions of patients. The MET call is similarly triggered by ward nurse identification of deteriorating patients, but in this model, doctors also attend the patient bedside. Unlike the reactive RRT and MET models, the PART and CCOT are proactive models that employ an Early Warning Score (EWS) or a track-and-trigger system.

In spite of some similarities and differences in the structure of all the outreach services, they all follow the common goal of shifting late identification/intervention to early identification/intervention (K. E. Davies, 2011). In a broader view, the Department of Health in the UK specifies three primary objectives of CCOT in hospitals: 1. to avert or prevent ICU admissions by early detection of deteriorating patients, 2. to support and monitor ICU discharged patients on wards and 3. to share critical care knowledge with staff in wards (DoH, 2000).

It was initially proposed that CCOT should be a multidisciplinary team led by a clinician expert in critical care (DoH, 2000). However, other compositions of outreach services were set up in hospitals, as well. For example, the RRT and MET are both medically-led teams as opposed to PART that is a nurse-led team. To shed light on the potential differences between the two structures, it is essential to conduct an in-depth analysis of their capabilities and efficiencies in improving patient outcomes. In addition, as the outreach team’s skills, experience and teamwork would significantly influence the timely and accurate detection and treatment of critically ill patients, there is still a paucity of research in the literature to illuminate the impact of these factors on patient outcomes. Risser et al. (1999) and Morey et al. (2002) elucidate that by improving the team’s skills and communication in an ED, medical errors noticeably decreased and the care quality improved significantly. In New Zealand also, in the empirical studies of Nurse Practitioners (NP) led by CCOT, Pirret (2008, 2012) found a steady decrease in the number of ICU readmissions within 72 hours after discharge and demonstrated that the NP are inclined to prescribe more medications
when they are given more authority. In the next section, we elaborate the elements that impact the performance of outreach services in hospitals.

1.5.1 The Team Composition and Size

One of the most influential factors that principally affect the efficacy of the outreach team in hospitals is the structure and size of the team (Cutler, 2006). Tobin and Santamaria (2012) discuss that the composition of the outreach team likely influences patient outcomes, as well. S. S. Scott and Elliott (2009) point out that the format of RRTs differs in various healthcare settings, and the particular implementation of such services depends on the available resources and staff, as well as the type of organisation where the RRT is aimed to be established. Similarly, Heintz and Schreiner (2007) and Arashin (2010) highlight that the size of the hospital, expected demands from different specialities and the availability of required resources such as ICU nurses and respiratory therapists determine the size of the RRT.

The findings of a survey carried out by Stolldorf and Jones (2015) on the size and structure of the RRT in the southeastern U.S. states also show that 68% of the RRT are staffed between 1 and 3 persons and the rest are constructed with 4 to 5 persons. It has also been reported that, while 94% and 95% of the RRTs are composed of ICU or ED RN and respiratory therapists, respectively, only 19% of hospitals have a dedicated RRT nurse who looks after patients in units and 3% of facilities have a dedicated RRT nurse who is not necessarily an ICU nurse. Besides, Figure 1.11 shows the characteristics of RRTs based on the type and size of the hospitals. It has also been observed that most large hospitals have more than two RRTs composed of an ICU or ED doctor and a devoted nurse. Therefore, it seems that the differences in size and type of hospital (e.g., teaching or non-teaching), as well as the availability of expertise and resources, are the primary reasons for the lack of universal similarity in optimum size (Schweickert, 2010) and structure (McDonnell et al., 2007) of the outreach services among hospitals.
Note. Hospital type: Teaching versus Non-teaching; Hospital size (number of beds licensed and staffed); S = 0-150; M = 151-500; L ≥ 500; Respiratory Therapist (RT); Rapid Response Team Member (RRTM)

Regarding the outreach structure, two distinct settings are commonly argued in the medical literature: nurse-led or medically-led (Cutler, 2006; A. L. Green & Williams, 2006). The PART and CCOT in most hospitals in the UK and New Zealand are a nurse-led team while the MET and RRT are a medically-led team established in most hospitals in the USA and Australia. Although clinical results reveal that there is an explicit disagreement amongst researchers on the functionality and productivity of the two structures, there is evidence that affirms the effectiveness of both approaches. Specifically, both streams of research that analyse the medically-led (Bellomo et al., 2004, 2003; Bristow et al., 2000; M. Buist, Harrison, Abaloz, & Van Dyke, 2007; M. D. Buist et al., 2002; DeVita et al., 2004; D. Jones et al., 2005; D. Jones, Egi, Bellomo, & Goldsmith, 2007; Kenward, Castle, Hodgetts & Shaikh, 2004) and nurse-led outreach teams find improvements in ICU performance and patient outcomes (e.g., ICU mortality admission rates) (Ball 2002; Esmonde et al., 2006; McDonnell et al., 2007; Story et al., 2006; Story, Shelton, Poustie, Colin-Thome, & McNicol, 2004).
1.5.2 Communication Strategy and Teamwork

A simple and efficient communication strategy, provides a beneficial tool for both ward staff and the outreach team to exchange information in a timely fashion (Cutler, 2006). Communication between both teams establishes a continuous intra-team collaboration (K. E. Davies, 2011). It is evident that a more efficient way of communication, liaison and collaboration among the medical and nurse staff plays a vital role in addressing the patients’ concerns (Aneman & Parr, 2006; Carter, 2008; D. Jones, George, Hart, Bellomo, & Martin, 2008; D. A. Jones, DeVita, & Bellomo, 2011; Priestley et al., 2004).

One of the applicable tools that assist both sides to enhance their communication level efficiently is the Early Warning Score (EWS) or track-and-trigger system. According to the definition of DoH (2013) “Early Warning Scores facilitate early detection of deterioration by categorising a patient’s illness severity and prompting nursing staff to request a medical review at specific trigger points utilising a structured communication tool while following a definitive escalation plan”. Although there are different versions of the EWS, most EWS are constituted of the physiological components of the patient (airway, respiratory rate/min, heart rate/min, systolic BP, conscious level, urine output and temperature) which are regularly measured and recorded by the ward nurses.

The national vital signs chart and the New Zealand Early Warning Score (NZEWS) are also designed for detecting clinical deterioration of acutely hospitalised, non-pregnant, adult patients (Health Quality and Safety Commission New Zealand, 2017). As shown in Figure 1.12, each patient’s vital signs are measured and documented. If any of these parameters deviates from the norm a score (0-3) is assigned to the patient. The score increases if the patient health condition deteriorates. The cumulative NZEWS and the total score will be finally obtained by adding the score of each vital sign. Subsequently, the appropriate action will be taken based on the escalation pathway (Figure 1.13). For example, if a patient is assigned an EWS 3, the nurse should then increase the frequency of monitoring and report the case to a senior nurse. In the most severe cases (EWS 10+) though, the immediate action of calling the RRT has to be placed by the ward nurses.
Note that, despite the significant implication of EWS in managing deteriorating patients being broadly discussed (Andrews & Waterman, 2005; Cuthbertson, Boroujerdi, McKie, Aucott, & Prescott, 2007; Goldhill & McNarry, 2004; H. Ryan, Cadman, & Hann, 2004; Subbe, Davies, Williams, Rutherford, & Gemmell, 2003; Subbe, Kruger, Rutherford, & Gemmel, 2001; Williams & Wheeler, 2009), there is still no common scoring method among hospitals (G. B. Smith, Prytherch, Schmidt, Featherstone, & Higgins, 2008), meaning that each hospital usually defines their own EWS (Gao et al., 2007; McDonnell et al., 2007; Peberdy et al., 2007). In summary, it is absolutely crucial to understand how efficiently the EWS should be designed and implemented in a hospital such that both ward and CCOT staff have a clear and common understanding of the illness severity of patients through an effective medical language.
1.5.3 Hours of Outreach Services

From the standpoint of hospital managers and policy makers, hours of outreach services is one of the most important factors during the implementation and evaluation phases of the outreach services (Moody & Griffiths, 2014). Some hospitals construct the outreach services along with the EWS pathway (Ball, 2002; Ball, Kirkby, & Williams, 2003; Esmonde et al., 2006; Fox & Rivers, 2001; McArthur-Rouse, 2001; Nassau, 2003; Subbe et al., 2003; Subbe, Williams, Fligelstone, & Gemmell, 2005) whereas others prefer to provide a telephone advice for 24/7 (McDonnell et al., 2007). In some settings also the outreach team visits the patients at the bedside or when they get discharged from the ICU (Ball, 2002; Ball et al., 2003; Esmonde et al., 2006; McDonnell et al., 2007) rather than being on call 24/7. Although the majority of the outreach services in both nurse-led and medically-led structures have been designed for 24/7, some hospitals cover the needs of their patients through a designated staff from critical care areas (Groom, 2001). Ideally, the service time should be 24/7 once the outreach team has been implemented, but it is rational to assume that the hospital setting
in general and the communication strategy and the team composition in particular, affect the amount of time the outreach service could function in a hospital. In particular, due to some economic and resource constraints, the service may not be available 24/7. However, it is necessary for the ward staff to know how and when they have to contact the outreach team (Cutler 2006).

1.5.4 Education and Team Skills

Education plays a pivotal role in detecting deteriorating patients (Chua, Mackey, Ng, & Liaw 2013; Cox, James, & Hunt 2006; Hart et al. 2014; McDonnell et al. 2013; Pantazopoulos et al. 2012). The skills and experience of the CCOT nurses could be thereby beneficial to junior staff in the ward in recognising deteriorating patients. Several works discuss the matter of insufficient skills and experience of junior ward staff in managing critically ill patients (M. Buist & Bellomo 2004; M. D. Buist et al. 2002; G. A. Harrison, Hillman, Fulde & Jacques 1999). Moody and Griffiths (2014) also elucidate that one of the leading factors that account for the suboptimal care of critically ill patients is lack of sufficient skill and experience of ward staff.

Different types of education can be provided by the CCOT to junior ward staff in the ward either at the bedside or as specific training programmes. It is noted that, an ongoing and extensive educational support supplied by the outreach team not only is the key factor of safely improving the health condition of severely ill patients (Bright, Walker, & Bion 2004), but also motivates the ward staff to be more involved in training programmes (A. L. Green & Williams 2006). In a survey study of 72 critical care nurse consultants performed in the UK, 48% of the nurses stated that they took the lead for developing training programmes outside the ICU/HDU (Dawson & McEwen 2006). M. Buist et al. (2007) also indicate that the performance of the MET can be improved through educational programmes. Finally, a recent prospective study carried out by D. Jones et al. (2006), indicated that through detailed training programmes and feedback support provided to
ward staff, the average number of MET calls per month increased significantly (from 25 calls per month to a maximum of 79 calls per month).

### 1.6 Aims and Objectives

A vast number of studies indicate significant association between nurse staffing and mortality in ICUs (e.g., Dimick, Swoboda, Pronovost, & Lipsett, 2001; Person et al., 2004; Pronovost et al., 2001). Nonetheless, proposing an ideal solution of employing more nurses to hospitals is practically infeasible due to two primary reasons: the costly expenses in recruiting nursing staff, and more importantly, the global shortage of nursing workforce. Regarding the latter, for example, the U.S. Bureau of Labor Statistics predicted a 15% growth in employment of registered nurses (RN) between 2016 and 2026, resulting in 438,100 vacancies in the healthcare system (Richards, 2012). In the UK, 90% of hospitals also reported a nursing shortage for December 2015 (Lintern, 2015). Similarly, a lack of 15,000 nurses is predicted in New Zealand hospitals by 2035 (Nanesh, Stokes, Molano, & Dixon, 2013). This concern is even more severe in ICUs as the nurse-to-patient ratios are much lower than those in other medical units such as a general ward.

Despite the nurse scarcity in the critical care setting, it is still vital for hospitals to come up with an effective and result-oriented policy to ensure the delivery of high-quality patient care. Establishing a costly nurse-led PART in hospitals was proposed by the policymakers aiming at improving the capacity of ICUs through preventing unplanned ICU admissions. However, the controversy in the results of clinical research on the patient outcomes after implementing PART in hospitals (Chaboyer, Foster, Kendall, & James, 2004; Chaboyer, Thalib, Alcorn, & Foster, 2007; Elliott et al., 2008; A. Green & Edmonds, 2004) induced the decision-makers to think of what nurse staffing between PART and ICU would result in the desired outcomes not only for the acutely ill patients but also for the hospital managers. This thesis, therefore, seeks to address this question and mainly presents two distinct but related practical approaches to hospital managers towards achieving an optimal nursing
policy between PART and ICU.

In Paper I, we develop econometric models to estimate the marginal values of a critical care nurse at ICU and at PART. We use those values along with ICU and PART nurses occupancy data to estimate the effect of moving one nurse from ICU to PART and vice versa on the patient hospital length of stay. In Paper II, we further propose a queueing model to examine the impact of different nursing between PART and ICU on the ICU patient mortality rate. The queueing theory would particularly assist the policy-makers to analyse patient flows between the general ward, PART and ICU and accordingly maintain the number of required nurses in each medical unit. As it is risky and often impossible to make an experiment with a real hospital, we built a discrete-event simulation (DES) model which is safer, cheaper and provides better insights into complex systems such as hospitals. Besides, the simulation model provides us with the ability to perform a ‘what-if’ analysis to gain a better understanding of the impact of different nurse allocations between PART and ICU on the ICU patient mortality.

In Paper I, we fit different models to the data, and in Paper II use the dataset to validate the simulation model employing a 12-month history (1 July 2015 to 30 June 2016) of 8,576 visits of the PART to 2,662 patients across all wards and post-ICU admission outcomes of patients admitted via referral from PART in one of the largest hospitals in New Zealand.
Chapter 2

Research Design and Methodology

2.1 Introduction

The purpose of this chapter is to discuss the research design and methodology applied in this thesis. Figure 2.1 presents explicitly the different steps taken in each paper to achieve the ultimate and overarching goal of proposing optimal nursing between ICU and PART in hospitals in this dissertation. Note that as Gauch, Jr., H.G. (2003) argued these steps are not constant and should be viewed as an iterative and dynamic process.

2.2 Research Design

Grove, Burns, and Gray (2012) define research design as “a blueprint for conducting a study with maximum control over factors that may interfere with the validity of the findings”. Parahoo (2006) describes research design as “a plan that describes how, when and where data are to be collected and analysed”. Polit-O’Hara and Beck (2006) define research design as “the researcher’s overall for answering the research question or testing the research hypothesis”. In this thesis, we applied the quantitative explanatory approach in Paper I and the quantitative exploratory design in Paper II. Specifically, we used the regression analysis in Paper I, and while maintaining the quantitative approach, we applied the analytical approach of queueing theory and discrete-event simulation in Paper II. On the
latter, although the simulation modelling was sometimes categorised under the predictive 
(e.g., weather prediction as in [Lenhard, 2005]) or explanatory (e.g., agent-based simulation 
as in [Axtell et al., 2002; Cederman, 2005; Dean et al., 2012; Sawyer, 2004; Tesfatsion, 2006]) 
research design, any policy formulation or decision making such as a cost-benefit analysis 
(Grüne-Yanoff & Weirich, 2010) or, as in our study, obtaining the optimal nursing between 
PART and ICU is counted as an exploratory design. In the following sections, we will 
discuss each research step in more detail.

2.2.1 Literature Review and Gap Analysis

Literature reviews should “objectively report the current knowledge on a topic“, “provide 
insight into the dynamics underlying the findings of other studies“, and also, they “may
offer more conclusive results than a single primary research study” (Green & Williams, 2006). Besides, reviewing previous works would assist researchers to identify gaps in the body of knowledge. There are several types of literature review (e.g., traditional narrative, systematic or meta-analysis) designed for different purposes and research methods (e.g., quantitative, qualitative or mixed-method) (see Grant & Booth, 2009). In this thesis, we specifically pursue the common traditional or narrative literature review style. According to Cronin, Ryan, and Coughlan (2008), this approach critiques a body of literature aiming at providing the reader with a more holistic background of the current knowledge as well as illustrating the importance of new study. It can also stimulate new research areas by spotting gaps or contradictions in a body of knowledge, thereby assisting the researcher to develop new ideas, research questions or hypotheses. This type of literature review also allows developing conceptual or theoretical frameworks (H. Cooper, Hedges, & Valentine, 2009; Coughlan, Cronin, & Ryan, 2007; P. Davies, 2000). Unlike meta-analysis, one of the significant benefits of narrative reviews is to cover all quantitative, qualitative, and mixed methodological approaches, while retaining credibility and precision (Rozas & Klein, 2010).

Firstly, we started by collecting and critically analysing journal papers, books and reports that discussed ICU operational related concerns and problems from different databases (such as JSTOR, PubMed, ScienceDirect, Web of Science, Wiley Online Library, and Google Scholar). In the next step, we classified the problems into different groups based on their distinct natures such as ICU patient flow optimisation and control (Dobson, Lee, & Pinker, 2010; L. V. Green, 2002; Griffiths, Price-Lloyd, Smithies, & Williams, 2006), ICU admission and discharge policies (Chan, Farias, Bambos, & Escobar, 2012; Kc & Terwiesch, 2011; J. Li, Dong, & Zhao, 2015), ICU staffing and workload (K. L. Brown, Pagel, Pienaar, & Utley, 2011; Duraiswamy, Welton, & Reisman, 1981; Griffiths et al., 2005; Hashimoto, Bell & Marshment, 1987), ICU bed capacity management (Barado et al., 2012; Lamiell, 1995; Masterson et al., 2004) and the critical care outreach services (CCOS) (Aneman & Parr, 2006; DoH, 2000; McArthur-Rouse, 2001; Pittard, 2003). Finally, we performed a gap analysis to find out the inconsistencies and issues that were not yet discussed or remained unsolved.
2.2.2 Designing Research Questions

The results of the previous step revealed that there is a significant gap in both OM and medical literature in demonstrating the efficacy of outreach services in hospitals. The clinical studies show equivocal results on the effectiveness of PART on ICU patient outcomes, ICU re-admission rate and ICU LOS. To measure the PART, RRT, CCOT or MET performance in hospitals, the majority of these studies applied a before-after study incorporating patient demographic variables (e.g., age, sex, ethnicity), APACHE severity scores and other medical variables. In this regard, one of the biggest studies conducted in Australia found that the introduction of the MET system to 23 hospitals did not significantly influence the incidence of cardiac arrest, unplanned ICU admission and unpredicted mortality (K. Hillman et al., 2005). On the contrary, other studies discovered that the establishment of RRT in hospitals resulted in reduction of hospital cardiac arrest (D. Jones et al., 2005; D. Jones, Bellomo, & DeVita, 2009; Maharaj, Raffaele, & Wendon, 2015; Winters et al., 2013), hospital mortality (J. Chen et al., 2014; D. A. Harrison et al., 2010; Maharaj et al., 2015), ICU re-admission rate (Niven, Bastos, & Stelfox, 2014) and adverse event (Bellomo et al., 2004). The controversial results observed in medical research initially motivated us to further investigate the role and functionality of the outreach services in hospitals.

Despite the interest in evaluating the performance of the outreach services in hospitals in the medical literature, there has been little discussion on this matter in the OM literature. Although papers studying different aspects of ICU management (e.g., Chan et al., 2012; Kc & Terwiesch, 2011; S.-H. Kim et al., 2014) acknowledge the potential impact CCOT can make on ICU performance, very little research has been done. Perhaps the work carried out recently by Gershengorn, Garland, and Gong (2015); Hu, Chan, Zubizarreta, and Escobar (2015) could be thought of as the most inclusive research in this area. Specifically, Hu et al. (2015) analysed the impact of proactively transferring patients to the ICU on patient outcomes and LOS. Unlike the common EWS system used by ward nurses to monitor the
health condition of ward patients, Hu et al. (2015) used a dynamic risk score of patients named an Early Detection of Impending Physiologic Deterioration score, version 2, (EDIP2) developed by Escobar et al. (2012) and Escobar, Gardner, Greene, Draper, and Kipnis (2013). The main distinguishing feature of EDIP2 is that it updates the severity score of patients every six hours throughout their journey in the ward. Their simulation results illustrate the significant effect of proactively transferring acutely ill patients to the ICU on reducing patient mortality and LOS. Gershengorn et al. (2015) also investigated the effects of adding a Physician Assistant (PA) to CCOT on patient hospital mortality and LOS via a retrospective study of two cohorts, one with adding the PA to CCOT and the other without altering the staff. Likewise, their findings show an improvement in hospital mortality, LOS and time-to-transfer to the ICU. As averting unplanned ICU admissions is one of the fundamental principles of the outreach services, it is arguable that there is still inadequate analytical analysis of the potential impact of PART on the ICU capacity. This particular gap in the OM literature was our second motivation.

Moreover, as discussed in Section 3.3.1, while hospitals have been struggling with the global nursing shortage, constructing a costly nurse-led PART constituting of highly skilled critical care nurses should be sufficiently justifiable for hospital managers. However, there is no exhaustive analysis of the efficiency of outreach services in hospitals taking both medical and operational variables into account. In other words, it is not yet well grounded if a hospital manager has decided to implement one of the outreach services in their hospital, what nurse staffing between PART and ICU would be more beneficial to both patients and the hospital. This gap in both medical and OM literature was our third motivation to pursue this study.

Hence, we undertook this study to shed new light on the nurse allocation problem between PART and ICU in a hospital from two different angles. In Paper I, we examine the impact of nurse occupancy in both PART and ICU on patient hospital length of stay considering both operational and medical variables. In Paper II, we model the dynamic of patient flow between ward, ED, and ICU to analyse the impact of different ICU nurse
staffing on ICU patient mortality rate. We believe that the approaches employed in this research and results obtained will contribute to both medical and OM literature towards a better understanding of critical care nursing between PART and ICU.

2.2.3 Identify Choice of Research Method

It is widely agreed that the choice of research method is mainly determined by the type of research questions or research problems (Blaxter, Hughes, & Tight, 2010; R. Kumar, 2010; Walliman, 2010). Numerous studies attempt to provide different classifications of research methods based on various criteria such as the nature of the problem (Blaxter et al., 2010), the type of data (Walliman, 2010) or the mode of enquiry (R. Kumar, 2010). However, there is a consensus among researchers that most studies can often be categorised as quantitative or qualitative, or even both.

“Quantitative methods emphasize objective measurements and the statistical, mathematical, or numerical analysis of data collected through polls, questionnaires, and surveys, or by manipulating pre-existing statistical data using computational techniques” (Babbie, 2010), “[they] focus on gathering numerical data and generalizing it across groups of people or to explain a particular phenomenon” (Muijs, 2010). On the contrary, qualitative research is “an inquiry process of understanding based on distinct and methodological traditions of inquiry that explore a social or a human problem. The researcher builds a complex, holistic picture, analyses words, reports detailed views of informants and conducts the study in a natural setting” (Srivastava & Thomson, 2009), “qualitative researchers are interested in understanding the meaning people have constructed, that is, how people make sense of their world and the experiences they have in the world” (Merriam, 2009). The research approach could also be an integration of both quantitative and qualitative methods (mixed method). According to Creswell and Clark (2011), mixed methods research “is a research design (or methodology) in which the researcher collects, analyses, and mixes (integrates or connects) both quantitative and qualitative data in a single study or a multiphase program of inquiry.”
Based on what we discussed in Section 2.2.2 and the definitions of different research approaches described above, it is evident that the quantitative method best suits our study. That is to say, first, we employ statistical and analytical techniques to model the nurse staffing problem between PART and ICU and subsequently, we use the de-identified secondary data supplied by the hospital to test our proposed models. Note that in the secondary data analysis, the data has already been collected by someone else for another reason (Johnston, 2014), but the dataset still has the potential to be used to address questions that are different from the primary study (Hodapp & Urbano, 2013).

As presented in Figure 2.1, we explicitly divided our choice of research method into two categories: quantitative and analytical. Note that although the analytical techniques used in Paper II can still fall under the umbrella of quantitative method, the nature of addressing the research questions is different from the statistical methods used in the quantitative approach in Paper I. In the paper II, we specifically use different mathematical techniques such as the queueing theory and simulation to analyse the performance of a complex and dynamic system such as ICUs. This goal would not have been achieved through statistical modelling techniques. Therefore, we first developed a queueing model of ward and ICU (Figure 5.6) to analyse the patient flow between these two medical units in the absence of PART. In this model, patients arrive to the ward at rate $\lambda_g$ and are served with service rate $\mu_1$. Patients might also be transferred to the ward from some other medical units such as ED or OR but in our model, we presume that all the patients arrive at the ward with the same arrival rate. The critically ill patients in the ward will be at some point sent to the ICU/HDU and served with service rate $\mu_2$. Upon staying in the ICU they are transferred back to the ward where they (most of them) finally recover. To investigate the role and efficacy of PART on ICU mortality, we proposed the PART queueing model (Figure 5.7). In this model, a single server (PART) with a service rate $\mu_3$, and two classes of high-risk and low-risk patients with arrival rates $r_1$ and $r_2$ respectively are contributed to the Ward-ICU queueing model. Through the simulation analysis, we have found out that constructing PART even with one ICU nurse reduce ICU deaths noticeably.
In addition, analytical approaches are particularly vital to understanding the performance of a system where designing different “what-if” questions would assist the researchers to choose the best scenario among a set of competing alternatives (Gokhale & Trivedi, 1998). For example, using the simulation approach, we were able to set up different nurse allocation scenarios between PART and ICU to obtain the best staffing policy under which the ICU mortality rate is the smallest. We will briefly discuss the statistical and analytical techniques used in this thesis in Chapter 3.

2.2.4 Ethics Approval

To examine the applications of our developed models in the real world, we contacted different hospitals in New Zealand. As a result, the ICU and PART teams in Middlemore Hospital were keen to provide us with available data and relevant information. Therefore, according to the Guiding Principles for Conducting Research with Human Participants at the University of Auckland, we developed our ethics application that allowed us to study both PART and ICU at Middlemore Hospital and also use the de-identified data supplied by the hospital. Our ethics application was approved by the University of Auckland Human Participants Ethics Committee (UAHPEC) on 11 March 2016 with the reference number: 016798.

2.2.5 Data Collection and Analysis

We were supplied with a 12-month history (1 July 2015 to 30 June 2016) of 8,576 visits of the PART to 2,662 patients across all wards and post-ICU admission outcomes of patients admitted via referral from PART. The data had been initially anonymised or de-identified by Middlemore Hospital so that individual data could not be linked to specific individuals. The data was mainly collected from two separated databases (PART and ICU). These databases are stored confidentially on secure servers in the hospital. The databases were linked by a unique patient identifier while still on hospital servers by researchers working within the hospital.
As the first step of processing the data, we had to confirm that the data was ‘clean’, meaning free from any inconsistencies and incompleteness (R. Kumar, 2010). Although initially, we encountered minor issues such as missing gender for a tiny number of patients, the quality of the data supplied by the hospital was acceptable for our analysis. For the missing data, we contacted the hospital a few times and requested them to address those cases with errors. Once the issue of data incompleteness was resolved, we have re-examined the data to make sure that there were no longer discrepancies. For Paper I, we specifically used R statistical software version 3.3.2 to validate and analyse the data, build the econometric models and interpret the outcomes. We applied a range of techniques that were already established and scripted in the R statistical environment to validate our data (for example, see Loo & Jonge, 2018; Teetor, 2011). For Paper II, we also used the results of our data analysis in R and particularly the Input Analyzer in Rockwell Arena simulation software version 14.5 for fitting the statistical distributions to the data.

Overall, our data can be categorised into two main types: demographics or patient-level data and operational data. The demographics data captures the characteristics of patients such as age, sex, illness severity score, while the operational data contains mostly data on patient flow across the hospital such as patient hospital admission and discharge date/time. Although Table 2.1 represents descriptions of all the data that we used in this thesis, we have specifically described some of the technical data as follows:

- **APACHE.2.Score**: The APACHE II score is a severity score calculated for ICU patients from variables taken from the first 24 hours of ICU admission (e.g. heart rate, respiratory rate, temperature etc.). It is an integer score from 0 to 71 representing increasing illness severity, and it has been well correlated with ICU risk of death (Knaus, Draper, Wagner, & Zimmerman, 1985). This fact is well-demonstrated in Figure 2.2 which shows that that the higher the total APACHE II score, the higher the risk of mortality.

- **PUP score**: PUP score is a Physiologically Unstable Patient (PUP) score composed of the six cardinal vital signs of a patient (respiratory rate, temperature, heart rate,
systolic blood pressure, level of consciousness and urine output). The PUP score is measured and ranked between 1 (lowest severity) and 5 (highest severity) for a ward patient. The PUP scores are frequently measured and recorded for each patient by ward nurses. If a patient gets PUP 1, the nurse has to increase the frequency of visits to two-hourly or even more if required. If the patient is assigned a PUP 2 to 4, the frequency of nurse visits will decrease to a half hour and the House Officer will be informed to review the patient’s documents within one hour. For any patient with PUP 5 or more or even a cardiac arrest, the ward nurses will place a MET or SET call. The visits would eventually end if it was realised that the patient has been deteriorating and is required to be sent to another unit such as the ICU. ICU discharged patients to the ward would be assessed similarly.

- **PART reason for visit**: There are different reasons that PART visits a patient in the ward including MET call, SET call, first visit, routine visit or other reasons.

- **PART episode**: One episode of PART is made up of several visits. Each episode is triggered by nurse/doctor referral, MET call, PUP score or ICU/HDU discharge events and terminated by patient death, palliative care, patient clinically improved or patient transferred to Emergency Care (EC), CCU or ICU/HDU.
• The total number of ICU nurses required in each shift: The total number of required nurses for each ICU shift was calculated based on the number of ICU patients in that shift as well as the illness severity of each patient. For example, a highly acute ICU patient might require two ICU nurses as opposed to an HDU patient who may need 1/2 capacity of an ICU nurse. The number of nurses needed for each ICU patient was computed by the hospital taking the illness severity of each patient into consideration. The total number of available ICU nurses in each shift was also given based on the actual number of nurses attending each shift. Note that the ICU nurses are rostered for day shift (07:00 - 19:00) and night shift (19:00 - 07:00) separately. It is important to emphasise that the nurse-to-patient ratio in the ICU is not necessarily 1:1 and is influenced mainly by the illness severity of the ICU patient.

• ICU source of admission: The data shows the number of patients transferred from ED, OT, ward, other ICUs in the hospital, other hospitals and ICUs in other hospitals to our current ICU. This data is specifically beneficial to our queueing and simulation models as it helps to characterise different patient flows between ward, ED, PART and ICU.

• Total number of patients in the hospital per day: This data is the count of patients in inpatient, Critical Care, Kidz, Neonates, Women’s Health and Mental Health - Acute and Mental Health wards at Middlemore Hospital at 7am. This data would help us to calculate the occupancy level of the hospital.
<table>
<thead>
<tr>
<th>Data</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>The age of a patient</td>
</tr>
<tr>
<td>Sex</td>
<td>The sex of a patient (Male/Female)</td>
</tr>
<tr>
<td>APACHE.2.Score</td>
<td>Acute Physiology and Chronic Health Evaluation (APACHE) II severity score for ICU patients</td>
</tr>
<tr>
<td>PUP score</td>
<td>Physiologically Unstable Patient (PUP) score</td>
</tr>
<tr>
<td>Hospital admission date/time</td>
<td>The data and time of patient admission to the hospital</td>
</tr>
<tr>
<td>Hospital discharge date/time</td>
<td>The data and time of patient discharge from the hospital</td>
</tr>
<tr>
<td>PART visit data/time</td>
<td>The date and time of PART visit to a patient</td>
</tr>
<tr>
<td>PART reason for visit</td>
<td>The reasons that the PART nurses visit a ward patient</td>
</tr>
<tr>
<td>PART duration of visit</td>
<td>The duration of PART visit to a patient</td>
</tr>
<tr>
<td>PART episode</td>
<td>One episode of PART is made up of several visits</td>
</tr>
<tr>
<td>ICU admission date/time</td>
<td>The data and time of patient admission to ICU</td>
</tr>
<tr>
<td>ICU discharge date/time</td>
<td>The data and time of patient discharge from ICU</td>
</tr>
<tr>
<td>Number of required and available ICU nurses</td>
<td>The total number of required and available nurses for each ICU shift.</td>
</tr>
<tr>
<td>ICU patient outcome</td>
<td>A binary variable representing the patient death/survival</td>
</tr>
<tr>
<td>ICU source of admission</td>
<td>Number of patients transferred from other medical units</td>
</tr>
<tr>
<td>Number of patients in the hospital per day</td>
<td>Count of patients in Middlemore Hospital inpatient beds at 7am.</td>
</tr>
</tbody>
</table>
At this point, we only described the raw data, which does not include operational variables such as ICU occupancy or PART occupancy and even performance variables such as ICU LOS. We provide a detailed description of the calculations we performed to derive these measures from the raw data in both papers accordingly. We will also present a statistical summary of the patient-level variables (e.g., age, sex and illness severity score) as well as a graphical description of some data to portray the relationship between different variables.

2.2.6 Verification and Validation of Proposed Models

Even though the concept of model accuracy is certainly comparative rather than definite, the decision-makers using the results of the proposed model are still concerned with whether the logic and specifications of the model are “correct” (Sargent, 2001). Indeed, it is particularly important to realise that a model would hardly represent precisely the system features. The concern of producing a sufficiently correct model though, could be addressed through the stage of model verification and validation. Broadly speaking, verification confirms the correctness of the model with respect to its conceptual model and assumptions (Banks, Carlson, Nelson, & Nicol, 2009; Carson, 2002), while validation checks if the model represents the real world via comparing the model performance metrics with their counterparts in the real system (Altiov & Melamed, 2001).

In the context of computer simulation, different techniques can be employed to verify a model such as reviewing the model by experts, creating a logic flow diagram, examining the model outputs for different reasonable inputs or using a debugger (Carson, 2002). Likewise, a model can be validated through comparing the model input-output transformations to corresponding input-output transformations for the real system (Naylor & Finger, 1967) using statistical testing such as t-test, event validity or face validity (Sargent, 2005). For example, in the event validity, the number of occurrences of a particular event in the simulation is compared with the number of correspondent events in the real system (a
patient death in a hospital can be considered as an event). The face validity is also similar to an expert elicitation where the expert is asked to approve the logic of the model.

We benefited from some of these techniques in our study. In the simulation study (Paper II), in addition to getting the ICU consultant’s opinion on the designed model, we explicitly used some of the features of the Arena simulation software to verify our queueing and simulation models. For example, in Arena, we displayed the current number in queue (NQ), the resource status (idle, busy), resource capacity (MR) or number busy (NR) of a resource to gain a better understanding of the dynamics of the queueing model. We were able to monitor and justify the occupancy levels of resources (nurses) in both PART and ICU through altering the arrival rate of patients. These features were notably helpful in finding out if any resources in ICU or PART were unexpectedly over-utilised or under-utilised. Another possibility that helped verify the logic of the model was to monitor the length of the queue in both PART and ICU under different scenarios (e.g., increasing the arrival rates while keeping the number of resources fixed to see if the queue was building up). Besides, we tested our model under extreme conditions such as having only one patient in the queue or using deterministic arrival times. As a result of different types of testing, we were reasonably confident that the proposed model was correct.

In the validation stage, we followed the approach of Altiok and Melamed (2001) and applied statistical hypothesis testing using the Student’s t-distribution test to compare the ICU and PART LOS between the simulation data and the real data provided by the hospital.

First, we took two samples of output data from the hospital (real-life) and the simulation model as follows:

1. A sample \( \{L_1, \ldots, L_N \} \) of observed averages of ICU LOS and PART LOS from the hospital over days \( i = 1, \ldots, N \)

2. A sample \( \{\hat{L}_1, \ldots, \hat{L}_N \} \) of estimated means of ICU LOS and PART LOS collected from runs of the simulation model over days \( i = 1, \ldots, N \)
Second, we defined a sample consisting of differences between the two above samples:

\[ G_i = L_1 - \hat{L}_1, i = 1, \ldots, N \]

which are approximately normally distributed with mean \( \mu_G \) and variance \( \sigma_G^2 \). Now we can form our hypothesis as follows:

\[
\begin{align*}
H_0: \mu_G &= 0 \\
H_1: \mu_G &\neq 0
\end{align*}
\]

Under the \( H_0 \), the statistic

\[ t_{N-1} = \frac{\bar{G} - \mu_G}{S_G/\sqrt{N}} \]

is distributed according to a Student’s t-distribution with \( N - 1 \) degrees of freedom, where \( \bar{G} \) and \( S_G \) are the sample mean and sample standard deviation of the sample \( \{G_1, \ldots, G_N\} \), respectively.

Third, for a significance level, \( \alpha \), the corresponding confidence interval satisfying

\[ P_r(t_1 \leq t_{N-1} \leq t_2) = 1 - \alpha \]

is equivalent to

\[ P_r(\bar{G} - t_{\alpha/2,N-1}S_G/\sqrt{N} \leq \mu_G \leq \bar{G} + t_{\alpha/2,N-1}S_G/\sqrt{N}) = 1 - \alpha \]

Where \( t_1 = t_{\alpha/2,N-1} \) and \( t_2 = t_{1-\alpha/2,N-1} \). As a result, \( H_0 \) cannot be rejected if the above confidence interval includes 0, demonstrating that the test supports model validity. On the contrary, the model will not be valid if the confidence interval contains 0.

### 2.2.7 Recommendation

The last stage of a research study is to draw conclusions from the obtained results and eventually make some inferences and recommendations to both academics and practitioners. It is generally perceived that final proposals are counted as the primary measure of evaluating the validity and quality of the research. It is therefore essential to discuss all
the research implications and provide evidence for each conclusion rather than presenting them as unjustifiable or obscure results. It is also substantial to illustrate the significance of the outcome and the delimitations of the study for future research. We will thoroughly cover this stage of the research in Chapter 6.
Chapter 3

Literature Review

3.1 Introduction

A review of the existing literature was carried out on academic books, journal articles, conference proceedings, technical reports and online websites related to the area of study.

The goal of this chapter is to discuss what work has been done on the subject and how our research filled the existing gaps in the body of knowledge. We divide this chapter into two broad sections. In the first section, we will particularly discuss the inconsistencies, gaps and contradictions identified by the Operations Research/Management Science (OR/MS) community on the ICU related topics argued in Chapter 1 such as ICU admission and discharge policies, ICU nurse shortage and ICU outreach services. In the second section, we will mainly review the literature on both quantitative and analytical methods applied in this study.
3.2 Intensive Care Unit in Operations Management

Literature

3.2.1 Background

In response to the polio epidemic, the first intensive care unit was set up at Kommunehospitalet in Copenhagen in 1953 (Reisner-Sénélar, 2009). Within two years, the idea of establishing ICUs in hospitals was acknowledged in the US as well (Grossman, 2004). A few years later, the first ICU in Australasia was opened in Auckland Hospital in New Zealand in 1958 and Vincent’s Hospital in Australia in 1961 (http://www.anzics.com.au).

Since the introduction of the first ICUs in the world, the domain of ICU services has remarkably extended, and now ICUs provide treatment to acutely ill patients with a broad range of life-threatening conditions caused by trauma, sepsis and other medical or surgical disorders. As ICUs started to expand, the number of clinical works in understanding and improving the complex mechanism of critical care services began to grow accordingly. ICU clinical studies are beyond the scope of this thesis, and we refer the readers interested in this literature to comprehensive reviews by Vincent et al. (2002) and Ristagno and Weil (2009). The discussion provided in the following sections predominately focuses on OR/MS studies and their contribution to understanding the ICU.

3.2.2 Modelling of Admission and Discharge Policies in ICU

3.2.2.1 ICU Performance and Patient Outcomes

On the subject of discharge policy in any health care unit, the work by Berk and Moinzadeh (1998) could be considered as the most primary and comprehensive discussion in the OM field. The paper principally investigates the different impact of discharge decisions on patient average LOS, system accessibility and average complication risk of discharged patients, especially when the health care unit faces resource shortage. To analyse the various aspects of discharge decisions, the proposed model also considers the utilisation rate of
the unit and the status of the patient, which is measured by the stage of recovery and the respective time that the patient has to spend in that stage. The results of their research indicate that an early discharge of patients improves the availability of the system even without negatively influencing patient health condition. The findings are also aligned with the results of medical papers (C.-L. Liu et al., 2017; Rosa et al., 2015).

Concerning ICU admission policies, Shmueli, Sprung, and Kaplan (2003) studied a queueing model to optimise the admission policy in the ICU, aimed at maximising the number of lives saved at the ICU. The paper analyses three different admission policies: First-Come-First-Served (FCFS), first-come-first-served for all the ICU patients whose expected incremental survival benefits exceed some hurdle (FCFS-H) and first-come-first-served for all the ICU patients whose expected incremental survival benefits exceed a bed specific hurdle (FCFS-BSH). Their findings show that although the FCFS policy maximises the ICU bed occupancy level, it ignores future patients who might benefit more from the ICU’s beds and thereby this policy cannot be optimal as opposed to FCFS-H which shows an almost 18% improvement in overall survival. Hosseinifard, Abbasi, and Minas (2014) employed stochastic dynamic programming to model the ICU discharge policies, as well. The central question of the paper is, how to choose and discharge the current premature patient (demand-driven discharge) from the ICU and admit a new critically ill patient to the ICU. Formulating the problem by integer programming and then using the simulation technique, their findings indicate that, when ICU LOS is non-memoryless, discharging patients with the lowest readmission risk is optimum. On the ICU admission decision, S.-H. Kim et al. (2014) discuss how the information provided to the ICU decision makers can influence the ICU’s performance and what changes in different scenarios could be possibly considered to improve the ICU admission decision process. One of the main distinguishing features of this paper is the analysis of the doctor’s behavioural aspects that influence the admission decision. The results illustrate how the patient outcome can be improved when the admission decision is made by both discretionary and objective criteria rather than only objective criteria. However, the results represent no noticeable influence
of ICU admission on mortality rate, and apparently, the ICU’s load has no considerable effect on the patient’s admission to the ICU.

Due to ICU beds constraint, some of the OT patients do not have a chance to get admitted to the ICU on time, and in the worst scenario, some can even be abandoned. In fact, the rejection of severely ill patients can sometimes result in a disaster. N. v. Dijk and Kortbeek (2008) and N. M. v. Dijk and Kortbeek (2009) studied the rejection probability for patients at both OT and ICU through the queueing model. They showed and numerically validated that the $M|G|c|c$ and $M|G|c-1|c-1$ system provided a secure lower and upper bound for rejection probability. J. Li et al. (2015) proposed a dynamic programming model to study the admission and premature discharge decisions in the ICU. Instead of using the FCFS admission policy, they have classified the ICU admission patients into two classes of acutely ill patients who cannot be discharged prematurely (class 1) versus those who are not as critically ill as the first group and can be discharged early (class 2). They found a threshold for the fixed number of available beds for the second class of patients. Their findings show that the request of a patient will be declined if the number of available beds is lower than the threshold; otherwise, the patient will be admitted to the ICU. S.-C. Kim and Horowitz (2002) investigated if using a daily quota system with a 1-week or 2-week scheduling window could improve ICU performance. Applying the simulation study, the findings illustrated that the hospital can significantly benefit from linking the scheduling of elective surgeries through a quota system to the ICU admission process. Kolker (2009) developed a simulation model to find out the maximum number of elective operations that can be scheduled to lower the number of ICU diversions due to ICU bed constraints. Although some additional daily surgeries should be bumped to other time slots, the results demonstrate that the ICU diversion can be removed entirely when maximum four elective operations are scheduled. Hagen, Jopling, Buchman, and Lee (2013) constructed a simulation model to examine three different ICU admission policies: smoothing the surgery schedule for elective patients, prioritising admissions by expected LOS or patient severity and early ICU discharge. The findings reveal that prioritising patients by illness severity
significantly reduced delays for acute patients, but also increased the average waiting time for all patients. Besides, they found that aggressive bumping remarkably increased both readmission and mortality rates.

In connection with the ICU admission criteria mentioned previously, SSCM (1999) also established a guideline for prioritisation of patients admitted to the ICU. In this model, patients are ranked based on their illness severity, those who would benefit most from the ICU care (priority 1) to those who would not benefit at all (priority 4):

- **Priority 1:** These are critically ill, unstable patients in need of intensive treatment and monitoring that cannot be provided outside of the ICU (e.g., post-operative or acute respiratory failure patients).

- **Priority 2:** These patients require intensive monitoring and may potentially need immediate intervention (chronic comorbid conditions who develop acute severe medical or surgical illness).

- **Priority 3:** These unstable patients are critically ill but have a reduced likelihood of recovery because of underlying disease or nature of their acute illness (e.g., patients with airway obstruction).

- **Priority 4:** These are patients who are generally not appropriate for ICU admission (e.g., patients with peripheral vascular surgery or patients who are permanently unconscious).”

In addition to the ICU admission and discharge policies, the problem of ICU readmission was also centred on in the OM literature. Chan et al. (2012) examined different ICU discharge policies, considering the ICU’s capacity and the patient readmission risk. Through an empirical study of patient flows in the ICU, the paper compares the impact of implementing each discharge policy on patient mortality rate and ICU readmission rate. The proposed discharge policy (greedy policy) shows that designing a predictive model of patient readmission risk that incorporates the discharge decisions with some index or
priority policies can improve the ICU’s throughput and in some operational regimes can even improve the mortality rate. The findings also show that the greedy policy performs better than the probability of re-admission index and the LOS index. Furthermore, the advantages and disadvantages of expediting the service rate over periods of congestion and, particularly when returns (re-admissions) are indispensable was discussed (Chan, Yom-Tov, & Escobar, 2014). To analyse how the speedup can affect the service time, the papers presented a state-dependent queueing model in which the service times and re-admission probabilities rely on the overloaded and underloaded states of the system.

Concerning service time in the ICU, Chan, Farias, and Escobar (2016) proposed a queueing model to examine the effect of boarding delay from the ED to the ICU on patient ICU LOS. Their results show that an increase in ED boarding times led to a longer ICU LOS. Kc and Terwiesch (2009) also investigated the effect of workload on service time and patient safety in two various healthcare delivery services. Their empirical results show that workers speed up the service rate when the workload increases, but this acceleration is not consistent, and long duration of increased overwork decreases the service rate in cardiothoracic surgery.

### 3.2.2.2 ICU Occupancy

Another significant area that drew the OM society’s attention is ICU bed capacity management. L. V. Green (2002) figured out that almost 90% of ICUs have not got adequate capacity for the patients. She proposed that the unit capacity should be planned based on service quality rather than an occupancy target. Ridge, Jones, Nielsen, and Shahani (1998) found that there is a nonlinear relationship between the number of beds, utilisation rate and the number of transferred patients. Costa et al. (2003) indicate that, due to nonlinearity and variability of factors estimating the LOS, the average number of patients expected in a year, mean LOS and a target utilisation cannot precisely predict the number of required beds. However, through simulation study and data analysis, they realised that models at the individual patient level including patient’s categories, arrival pattern and LOS estimate
the number of ICU beds more accurately. Troy and Rosenberg (2009) and Zilm and Hollis (1983) also applied simulation study to obtain the required number of ICU beds for surgery patients. Cahill and Render (1999) analysed the ICU bed availability in Cincinnati VA Medical Center through the simulation study. They realised that having more telemetry beds, transferring ICU patients requiring the ventilator to a Respiratory Care Unit (RCU) and assigning ICU swing beds to the emergency room area could result in a higher accessibility of ICU beds. Applying the same analytical approach, Shahani, Ridley, Nielsen and Intensive Care Society’s and Department of Health’s Working Group on Patient Flows (2008) found that increasing ICU capacity can considerably reduce deferral and transfer rates.

3.2.3 Modelling of Uncertainties in ICU

3.2.3.1 Classifying Patients

Speaking of the heterogeneity of patients arriving at the ICU, previous works mainly divided the patients into two main categories of emergency or elective patients (Costa et al., 2003; Dobson et al., 2010; Griffiths et al., 2005; Hagen et al., 2013; McManus et al., 2004; Ridge et al., 1998). Cahill and Render (1999) and Akkerman and Knip (2004) classify patients into inpatients and outpatients. S.-C. Kim, Horowitz, Young, and Buckley (1999) and S.-C. Kim and Horowitz (2002) took even one step further and broke down the inpatients into wards and OT patients (emergency or elective). N. v. Dijk and Kortbeek (2008) and N. M. v. Dijk and Kortbeek (2009) categorised patients based on the OT’s referral. Kolker (2009) grouped ICU patients based on the referrals received from external hospitals. Barado et al. (2012), Asaduzzaman, Chaussalet, and Robertson (2010), Adeyemi, Demir Chahed, and Chaussalet (2010) and Cochran and Bharti (2006a) classified patients based on the ICU referral departments (e.g., OT, ED, wards etc.). Some patients are also classified based on their illness severity (Chan et al., 2012; J. Li et al., 2015; Wharton, 1996). Our approach in Paper II is a combination of the last two streams: we model the ICU referral
patients from ED and ward departments as well as classify the patients based on their severity, while in Paper I, we only take the illness severity of patients into account.

### 3.2.3.2 Arrival Process and Length of Stay

The ICU arrival process and LOS are also modelled through different statistical distributions. The majority of works use the Poisson distribution as the arrival pattern to the ICU (Asaduzzaman et al., 2010; Cahill & Render, 1999; L. V. Green, 2002; Griffiths et al., 2006; Hagen et al., 2013; Kolker, 2009; J. Li et al., 2015; Litvak, van Rijsbergen, Boucherie, & van Houdenhoven, 2008a; McManus et al., 2004; Shmueli et al., 2003) while Dobson et al. (2010); Griffiths et al. (2005); S.-C. Kim and Horowitz (2002); Masterson et al. (2004) and Lowery (1992) apply the non-stationary Poisson distribution instead due to the changes in the ICU arrival process within a day. Studies using the Markov Decision Process (MDP) and queueing theory normally modelled the ICU LOS as an exponential or geometric distribution because of its memoryless property (Asaduzzaman et al., 2010; Chan et al., 2012; Cochran & Bharti, 2006b; Dobson et al., 2010; L. V. Green, 2002; S.-H. Kim et al., 2014; J. Li et al., 2015; Litvak et al., 2008a; McManus et al., 2004; Shmueli et al., 2003). Papers applying regression models the ICU LOS as an exponential distribution (Adeyemi et al., 2010; Shmueli, Baras, & Sprung, 2004; Shmueli & Sprung, 2005), as well. On the contrary, Kc and Terwiesch (2011) fit the Weibull distribution to the ICU LOS and discuss that the memoryless property is not realistic as the service speeds up the process when its get highly loaded. Other papers model the ICU LOS as a lognormal distribution (Chan et al., 2012; Masterson et al., 2004; Shahani et al., 2008), triangle distributions (Cahill & Render, 1999) or phase type distributions (Demir, Lebcir, & Adeyemi, 2014). The ICU interarrival times and LOS are modelled as the exponential and lognormal distributions, respectively, in Paper II. Similar to S.-H. Kim et al. (2014) and other papers mentioned above, we fit the lognormal distribution to our ICU LOS and HOSLOS in our econometric models in Paper I.
3.3 Nurse Staffing and Rostering

3.3.1 Background

The literature on nurse personneling can be divided into two main areas: nurse staffing and nurse rostering. The majority of works focused on the **Nurse Rostering Problem (NRP)** or the **Nurse Scheduling Problem (NSP)** to optimise the allocation of nurses to different time slots. The works on nurse staffing, however, struggled to obtain the optimum number of nurses in hospital units.

3.3.2 Nurse Rostering Problem

In NRP, mathematical programming is commonly used as an optimisation approach. Minimising staff requirements and maximising individual needs and preferences are considered as some of the goals in these approaches (Cheang, Li, Lim, & Rodrigues, 2003). Baker (1974) proposed a cyclic scheduling to assign two consecutive days off in each week to full-time staff, aiming at minimising staff size. To minimise the cost of the workforce, Bartholdi, Orlin, and Ratliff (1980) considered the cycling staffing problem using an integer programming model. A cyclic scheduling approach has also been applied by Burns and Koop (1987), considering two days-off each week with three types of work shift. In this arena, however, non-cyclic models were also taken into account. On this subject, Warner and Prawda (1972) constructed a mixed-integer quadratic programming model to minimise the “shortage cost” of nursing over a period of 3 to 4 days. Nursing skills class, nursing personnel capacity, integral assignment, nursing shifts and other constraints were considered in the model. Warner (1976) further constituted a large multiple-choice programming model to maximise nurse’s preferences, regarding the length of work stretch, rotation patterns and request for days off. Miller, Pierskalla, and Rath (1976) modelled NSP that minimises an objective function that balances the trade-off between staffing coverage and schedule preferences of individual nurses. Millar and Kiragu (1998) also presented a shortest-path problem with side constraints, for cyclic and non-cyclic nurse
scheduling with two work shift types. The paper proposed a new pattern named a stint which considers a start date, a length, a cost and the shifts worked.

Specifically on the ICU, due to the high level of uncertainty, the modelling of ICU nurse staffing is complex; however, the number of works devoted to this area is not as comparable as those in other hospital’s departments (Bai et al., 2018). Hashimoto et al. (1987) simulated a 12-bed medical/cardiac ICU workload and staffing system. They considered the nursing policies, costs, and availabilities, and patient illness severity per shift as the input and estimated the total overstaffing, understaffing, and cost per year for FTEs nursing for a direct patient. Their results reveal that the optimum nursing would be based on 5.5 direct FTEs per shift. In Section 3.3.1, we argued that the availability of ICU nurses and not ICU beds largely impacts the ICU admission decision. Griffiths et al. (2005) also illustrated this fact and assumed that although the number of ICU beds can be increased, the number of ICU nurses is constant. They applied a simulation model to obtain the required number of supplementary nurses per shift considering the time-dependent nature of elective surgery admissions and different ICU LOS profiles. Besides, the study argues the effect of implementing a nurse-led outreach service on the ICU LOS as well as the ICU total costs. The simulation results show that if the implementation of the outreach team could reduce the ICU LOS by 20%, the ICU would then need one less nurse per shift thereby reducing the average cost per shift by 0.51X where X is the cost of employing a rostered nurse. Unlike other works in ICU nursing, Duraiswamy et al. (1981) looked at the shift scheduling problem in the ICU and used the simulation study to estimate the required number of nurses for ICU based on the nurse occupancy level, patient arrival patterns and illness severity. Mullinax and Lawley (2002) studied the problem of large variation in infant conditions in a Neonatal Intensive Care Unit (NICU) and its impact on the ICU nurse workloads. They proposed an integer linear programme that assigns infants with different severities to nurses while balancing nurse workloads.
### 3.3.3 Nurse Scheduling Problem

The literature on NSP also shows that limited research uses a queueing approach to model nurse staffing. The efficacy of nurse-to-patient ratio mandated by California Bill AB 394 was examined by [de Véricourt and Jennings (2008)](https://journals.sagepub.com/doi/10.1177/0007112506292626). They considered that there is a meaningful relationship between delays in meeting patient requirements and adverse medical outcomes and in order to analyse the performance of this regulation in the queueing model, the frequency of excessive delay has been considered as a metric. They present a finite population $M/M/s//n$ queue where the patients have two states: needy or stable. The paper develops two heuristic staffing policies (efficiency-driven (ED) and quality-and efficiency-driven (QED) staffing regimes) as the solution to the queueing model. Their findings demonstrate that, depending on the size of the medical units (small or large), the quality of care varies significantly. In another work, [de Véricourt and Jennings (2011)](https://journals.sagepub.com/doi/10.1177/0007112511399357), by considering the workload experienced by nurses, formulated the number of required nurses in medical units, using a closed $M/M/s//n$ queueing system where $s$ is the number of homogeneous nurses within a single medical unit and $n$ is the number of homogeneous patients. To obtain the number of nurses, they also proposed a heuristic result in a new staffing regime. Similar to their previous study, the probability of excessive delay indicates the performance of the medical units. Their findings illustrate the possible relationship between the frequency of excessive delay and its impact on patient outcome. [Yankovic and Green (2011)](https://journals.sagepub.com/doi/10.1177/0007112511399357) developed a finite source queueing model with two sets of servers (nurses and beds) to estimate the actual interdependent dynamics of bed occupancy levels and demands of nursing. They also examined how unit size, nursing intensity, occupancy levels and unit length-of-stay influence the nurse-to-patient ratio and nurse efficiency in ED. [L. V. Green, Savin, and Savva (2013)](https://journals.sagepub.com/doi/10.1177/0007112511399357) developed a single-period newsvendor model to analyse the impact of nurse absenteeism on nurse staffing. Their findings indicate that absentee rates are consistent with nurses showing an aversion to higher levels of expected workload and the rate of nurse absenteeism would decrease when the hospital unit planned to schedule an extra nurse.
Heuristic and meta-heuristic methods were also applied as a solution approach to NRP. Tabu Search (TS) (E. Burke, Causmaecker, & Berghe, 1998; Dowsland, 1998; Nonobe & Ibaraki, 1998), Genetic Algorithm (GA) (Aickelin, 2010a; Aickelin & Dowsland, 2004; Aickelin & White, 2004; Beddoe, Petrovic, & Li, 2009), branch-and-price (Beliën & Demeulmeester, 2008), and Simulated Annealing (SA) (Thompson, 1996) are some of the works that employ these algorithms to solve NRP. With respect to decision variables applied in NRP or NSP, three main patterns can be seen: 1. daily pattern (Abdennadher & Schlenker, 1999; Balakrishnan & Wong, 1990; Bard & Purnomo, 2005; Jaumard, Semet, & Vovor, 1998; Millar & Kiragu, 1998), 2. shift pattern (Aickelin, 2010b; Dowsland, 1998; Maier-Rothe & Wolfe, 1973) and 3. task pattern (E. Burke, Cowling, Causmaecker, & Berghe, 2001; Jaumard et al., 1998; Warner, 1976). Another aspect of mathematical modelling of NRP widely discussed in the literature is the constraints that can be generally sorted into two classes of hard and soft constraints. Hard constraints commonly encompass coverage requirements such as nurse demand per day per shift per skill class, whereas soft constraints mainly consider the time requirement on nurse scheduling (Cheang et al., 2003). Nurse workload (Al-Yakoob & Sherali, 2007; Burns & Koop, 1987; Kostreva & Jennings, 1991; Musa & Saxena, 1984), working shift/days (Burns & Koop, 1987; Kostreva & Jennings, 1991; Sitompul & Randhawa, 1990), nurse skill (Harper, Powell, & Williams, 2010; Warner & Prawda, 1972), nurse preference (Bard & Purnomo, 2005; Berrada, Ferland, & Michelon, 1996; De Grano, Medeiros, & Eitel, 2009; Kostreva & Jennings, 1991; Valouxis & Housos, 2000; Warner & Prawda, 1972), shift types (E. Burke et al., 2001; Valouxis & Housos, 2000), working weekends (Burns & Koop, 1987; Musa & Saxena, 1984; Valouxis & Housos, 2000; Ásgeirsson, 2014), shift patterns (E. Burke et al., 2001; Ásgeirsson, 2014) and demand for different types of nurses (Burns & Koop, 1987; Kostreva & Jennings, 1991; Musa & Saxena, 1984) are some of constraints that have been used in NRP (for more details, see E. K. Burke, Causmaecker, Berghe, & Landeghem, 2004; Cheang et al., 2003; van Dam, Meewis, & van der Heijden, 2013).
3.4 Rapid Response Systems

3.4.1 Introduction

It is widely argued that patients in the ward often develop some signs of health deterioration a few hours before cardiac arrest (M. D. Buist et al., 1999; Franklin & Mathew, 1994). However, early identification and treatment of such potential high-risk patients might prevent progression to cardiopulmonary arrest (Bedell, Deitz, Leeman, & Delbanco, 1991; Schein, Hazday, Pena, Ruben, & Sprung, 1990; A. F. Smith & Wood, 1998). It is even observed that a noticeable percentage of severe adverse events (approximately 20%) are preventable (Neale, Woloshynowych, & Vincent, 2001). To address the “failure to rescue,” the Rapid Response Systems (RRS) were proposed to hospitals (D. A. Jones et al., 2011). Specifically, the RRSs aimed to identify the acutely ill patients in the ward as promptly as possible and accordingly treat and monitor them based on a care plan (Devita et al., 2006).

As presented in Figure 3.1, the RRS constitutes two afferent and efferent limbs as well as the administrative and quality improvement functions (Devita et al., 2006).

Figure 3.1 Rapid Response System Structure (Source: Devita et al., 2006)
3.4.2 Afferent Limb

The afferent limb is one of the critical elements of the RRS that detects an event and triggers responses to urgent needs of a patient. The standard activation criteria for the RRS are primarily based on the airway/breathing, circulation and neurological symptoms (K. Hillman et al., 2005). In addition to these objective criteria, Santino et al. (2009) observed that the instinct feeling of “nurse worried” is the most common subjective criterion that triggers a call for RRT. They realised that more than 50% of “worried” calls can be justified with one of the objective criteria. To design an efficient afferent limb, three components required to be taken into consideration: 1. a set of sensitive and quantifiable measures, 2. a human and technological track and trigger system and 3. a communication strategy (Moore, Hravnak & Pinsky, 2012).

The afferent limb though could fail which may result in patient death (Boniatti et al., 2014; Quach et al., 2008; Trinkle & Flabouris, 2011). A suggested solution to address Afferent Limb Failure (ALF) is to use an automated EWS. In the US, for example, the application of algorithm-based-automated surveillance systems along with electronic interfaces is growing quickly (Hravnak et al., 2008; Rothman, Solinger, Rothman, & Finlay, 2012; Zimlichman et al., 2012). One example of an automated EWS is Visensia® from OBS Medical. Visensia’s Artificial Intelligence continuously measures the vital signs of a patient (heart rate, respiration rate, blood oxygen saturation, temperature and blood pressure) and compares the outcome with a set of real word patient dataset to detect any pattern of instability in the patient health. The measured vital signs are later fused into one objective numerical index and scaled from 0 to 5 with three categories of 0 = Normal, 3 = Visensia Alert (caregiver attention necessary) and 5 = Critical (Figure 3.2). Eventually, Visensia generates a response such as an alarm for the critical patients accordingly.

3.4.3 Efferent Limb

All the MET, RRT, PART and CCOT that we discussed previously are different models of a RRS that construct the efferent limb. Some of these structures are nurse-led RRS such
as PART while RRT and MED teams are mostly medically-led RRS, the CCOT in the UK is commonly staffed with the ICU trained nurses working in close collaboration with the ICU consultants. Although these approaches differ in size and composition, they should all be able to identify the high-risk patients in wards and exert an appropriate level of intervention accordingly. Typically, the medically-led “high capability” teams as opposed to a nurse-led “intermediate capability” or “ramp-up” teams (Devita et al., 2006) are able to provide a broader range of treatment and services to needy patients (Garcea, Thomasset, McClelland, Leslie, & Berry, 2004). However, the professional nurse-led team could be more cost-effective to a hospital. Table 3.1 compares the strengths and weaknesses of the two structures in more detail.
Table 3.1 The Advantages and Disadvantages of two Effector Limb Response Team Models (Source: Devita et al., 2006)

<table>
<thead>
<tr>
<th>Model effect</th>
<th>High capability</th>
<th>Intermediate capability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advantages</td>
<td>Definitive care provided as quickly as possible</td>
<td>May feel less intimidating, resulting in earlier calls</td>
</tr>
<tr>
<td></td>
<td>“One-stop shopping” for services</td>
<td>May be less expensive</td>
</tr>
<tr>
<td>Disadvantages</td>
<td>May be intimidating to nurses, leading to delayed calls</td>
<td>May be less efficient, resulting in a delay of care</td>
</tr>
<tr>
<td></td>
<td>Requires highly trained (first) responders; may be costly</td>
<td></td>
</tr>
</tbody>
</table>

3.4.4 Administrative and Quality Improvements

Lastly, administrative and quality improvements mainly look after the planning, implementation and maintenance of equipment, data collection and analysis, quality improvement and patient safety during different phases of RRS (Devita et al., 2006). The constructive role of preventive maintenance is mainly known in crisis particularly when the patient, for example, needs a defibrillator, but it is out of service due to unexpected malfunction.

3.5 Queueing Theory

3.5.1 Origins, Standard Results and Applications

Queueing theory was first introduced by Erlang in 1904 to identify the capacity required in the Danish telephone system (L. Green, 2006). Since then, queueing models have been applied in many service industries such as banks, airlines and call centres (e.g., Brewton, 1989; Brigandi, Dargon, Sheehan, & Spencer, 1994; Brusco, Jacobs, Bongiorno, Lyons, & Tang, 1995; Holloran & Byrn, 1986; Stern & Hersh, 1980), and emergency systems such as police patrol and fire and ambulance (e.g., Chelst & Barlach, 1981; L. Green & Kolesar, 1984; Kolesar, Rider, Crabill, & Walker, 1975; Larson, 1972; Taylor & Huxley, 1989). The applications of queueing theory, however, are not only limited to these areas. As illustrated
in Figure 3.3, queueing theory has also been widely used in healthcare. Nevertheless, the main purpose of this section is to briefly point out the applications of queueing models in healthcare that have not been discussed in previous sections, such as the ambulatory services or the pharmacy services. For more details on this subject, we refer the readers to Lakshmi and Iyer (2013).

One of the most critical factors in managing medical emergency cases is to provide the patient with the ambulance services as quickly as possible (Lakshmi & Iyer, 2013). The number of ambulances required to achieve particular response rates (Bell & Allen, 1969), the average response time of the system to an emergency call in an ED (Mendonça & Morabito, 2001) and the prediction of the whole distribution of the response time, particularly the rate and spatial distribution of demand, variable ambulance velocities and queueing effects (D. W. Scott, Factor, & Gorry, 1978) were the problems that modelled via the queueing theory.

Figure 3.3 Application of Queueing Theory in Health care (Source: Lakshmi & Iyer, 2013)
Queueing theory has also been used in the pharmacy services. Prescription queues and work measurement of prescription fill times (Donehew & Hammerness, 1978), modelling the waiting times of prescriptions with different priorities (Shimshak, Gropp Damico & Burden, 1981), patient waiting time in the outpatient pharmacy (Vemuri, 1984), and analysing the relationships between the number of pharmacy staff members, the prescription dispensing process and outpatient waiting time (Moss, 1987) are the main problems that were discussed in this area.

Queueing theory is the mathematical study of waiting lines, or queues (Sundarapandian, 2009). A queueing model provides ways to estimate the length and waiting time in a queue (Sundarapandian, 2009). As depicted in Figure 3.4, a queueing system is characterised by the arrival process, service time and service discipline (Prabhu, 1997) and represented by the Kendall’s notation in the form of A/S/c/K/N/D (Kendall, 1953) where,

- \(A\): the interarrival time distribution
- \(S\): the service time distribution
- \(c\): the number of servers
- \(K\): the system capacity (the maximum number of customers in the system)
- \(N\): the population size
- \(D\): the service discipline

![Figure 3.4 A Queueing System](image.png)

To analyse a queueing model, we require information on the arrival process. Specifically, we should be able to identify how customers arrive at the system (single or batch), what
is the distribution of interarrival times between the arriving customers and the size of population (finite or infinite). Besides, we need to understand the service mechanism including the number of servers, the capacity of the system, the service time distribution and the service discipline. The service discipline establishes rules on how customers get selected for the service. Overall, there are four common queueing disciplines as follows (Sztrik, 2016):

- FIFO: First-In-First-Out: who arrives earlier leaves earlier
- LIFO: Last-In-First-Out: Who arrives last leaves earlier
- RS: Random Select: a customer gets selected randomly, and
- Priority: a customer gets picked based on different priority rules (e.g., in an ED, critical patients get triaged first)

Finally, as Prabhu (1997) described, we would need to find out if the queueing system has:

- Balking: Customers do not join the queue as it is too long
- Reneging: Customers join the queue but leave it later on as they have waited too long, or
- Jockeying: Customers switch between queues as they think they get served faster in other queues

The key variables used in the queueing models are represented as follows:

- $\lambda$: Mean arrival rate
- $W$: Average waiting time in a system
- $W_q$: Average waiting time in a queue
- $L$: Average number of jobs in the system
• $L_q$: Average number of jobs waiting in a queue

• $S$: Service time per job

• $\mu$: Mean service rate per server = $1/E[S]$

• $\rho$: The utilisation of server (the probability that the server is busy), hence we have the following equation:

$$W = W_q + E[S]$$  \hspace{1cm} (3.1)

In addition, according to Little’s law, the long-term average number of customers ($L$) in a stationary system (the joint probability distribution, mean and variance remain unchanged over time) is equal to the long-term average effective arrival rate ($\lambda$) multiplied by the average time that a customer spends in the system ($W$) (A. O. Allen & Rheinboldt [2014]; Leon-Garcia [2008]):

$$L = \lambda W$$  \hspace{1cm} (3.2)

Therefore, based on the Little’s law, we can obtain the number of jobs in the queue as well as the utilisation of the server:

$$L_q = \lambda W_q$$  \hspace{1cm} (3.3)

$$\rho = \lambda / \mu$$  \hspace{1cm} (3.4)

The complexity of a queueing model, however, can arise when the arrival and service processes follow different probability distributions or there is more than one server in the system. For example, the $M/M/c/K$ queue is a multiserver, finite-capacity queue where the arrival rates follow the Poisson distribution and the service times are exponentially distributed. The number of jobs in the system is a birth-death process which is a special type of continuous-time Markov process in that the state of the system at one point in time
sometimes increases only by one, and at other times decreases only by one (Zukerman, 2013).

With respect to the queueing networks, the product-form solution, under certain assumptions, allow one to acquire a simple exact solution for the joint queue length distribution in a separable form (N. M. V. Dijk, 1993; D. R. Smith, 1983; Walrand, 1988). In other words, they provide a simple closed-form expression of the stationary state distribution that allows defining efficient algorithms to evaluate average performance measures and specific characteristics of the system such as state-dependent routing and departures and finite capacity queues (Balsamo, 2000). However, not all queueing systems have a product-form solution which requires applying other solution approaches such as approximations or a simulation. We will further discuss the main solution approaches of fluid and diffusion approximations and discrete-event simulation that are mainly applied by the previous studies.

3.5.2 Intractability and Practical Applications

Once analytical solutions for the queueing models become intractable, approximation techniques can be applied as alternative approaches (Boxma, Koole, & Liu, 1994). For example, open queueing networks, nonstationary queues, or queues with non-exponential interarrival and service times usually do not provide a closed-form solution. Therefore, to study the behaviours of the queue and most importantly obtain key results such as the length of the queue or waiting time in the queue, one could apply diffusion and/or fluid approximations. In Sections 3.5.2.1 and 3.5.2.2, we will briefly discuss some of these approaches towards addressing the intractability in the queueing models.

3.5.2.1 Heavy and Light Traffic Approximations

A heavy traffic or diffusion approximation is a technique in which an appropriate diffusion process (Heyman & Sobel, 1990) replaces a complicated and analytically intractable stochastic process. Heavy traffic limit theorems approximate the performance measures of
the system when it is nonempty and operates near its full capacity, whereas light traffic theorems obtain the desired performance measures when the system is very lightly loaded (Varma, 1990). Reflected Brownian motion (RBM) was also proposed by Kingman (1962) then approved by Iglehart and Whitt (1970) to characterise queueing models with heavy traffic. In heavy traffic, for instance, the queue length can be approximated by reflected Brownian motions with drift. The light traffic approximation was also developed for a single queue (Burman & Smith, 1983, 1986) and an open queueing network (Reiman & Simon, 1989).

Another useful approximation that was primarily used to solve the intractable queueing model is a fluid approximation. In a fluid model, a divisible commodity (fluid) arrives at a storage facility where it is stored in a buffer and gradually released (Whitt, 2002). Consider a standard queueing model in which individual customers or jobs are arriving at the service centre, probably wait for the service, are served and then leave the system. For such queueing models we compute the number of customers in both the queue and the entire system and explain the experience of individual customers (Whitt, 2002), such as their average waiting times in the queue. Conversely, a fluid queue model is applied to the cases in which the individual customer is so small; thereby instead of considering the discrete behaviour of individuals, we can consider them as a continuous stream of work that flows into the system (Arunachalam, Gupta, & Dharmaraja, 2010).

### 3.5.2.2 Large Deviations Theory

The theory of large deviations is concerned with the probabilities of rare events that are exponentially small as a function of some parameter (Touchette, 2011). The applications of large deviations theory can be observed in various problems where detailed information on a rare event is needed (Ellis, 2008). The analysis of the tails of a probability distribution, for example, can be considered as one of its applications in queueing theory (Lewis & Russell, 1997). This technique was also used in both single queue (Chang, 1994; Duffield, O&#39, &
3.5.2.3 Simulation

Another alternative solution to complex queueing models without closed-form solutions is a computer simulation. Due to the programmability of the simulation, any queueing model in the real world can be duplicated with a high level of accuracy and detail in the simulation model. This could be counted as the main reason why simulation models have become a very common and favoured method of analysis. According to Shannon (1975), "simulation is the process of designing a model of a real system and conducting experiments with this model for the purpose either of understanding the behaviour of the system or of evaluating various strategies (within the limits imposed by a criterion or set of criteria) for the operation of the system." Simulation is the imitation of the operation of a real-world processor system over time (Banks, Carson, Nelson, & Nicol, 2010). It is mostly employed 1. to understand and interpret different behaviours of a complex system, 2. to help develop hypotheses and theories based on the observed behaviour of the system and 3. to predict the future behaviour of the real system (Fàbrega, Vilà, Careglio, & Papadimitriou, 2013). The simulation model is mainly useful when the mathematical model is too complex to be solved by analytical methods (Fishman, 2001). In fact, the only way to solve the mathematical equitations is numerically using computer software; this process is called a computer simulation (Sokolowski & Banks, 2010).

Note that modelling of a very complicated system is not the only advantage of simulation. Comparing new policies without disrupting the real operation, performing "what-if" and bottleneck analysis, examining the behaviour of a system when it is not built yet or safe to be experimented (Banks et al., 2010), observing the detailed structure and overall behaviour of the system by manipulating time (expanding or compressing time), controlling the source of variation, resuming and restoring the state of the system (Fishman, 2001) and flexibility and cost-effectiveness (Kelton, Sadowski, & Zupick, 2014) are counted as other
benefits of simulation study. These are the important features of simulation modelling that encourage both policymakers and scholars in the healthcare industry to apply simulation in different areas such as managing patient flow and staff scheduling (Dittus, Klein, DeBrota, Dame, & Fitzgerald, 1996; Ingolfsson, Erkut, & Budge, 2003; Vasilakis, Sobolev, Kuramoto, & Levy, 2007), improving service outcomes (Duguay & Chetouane, 2007; Harper, Shahani, Gallagher, & Bowie, 2005; Stahl et al., 2004; Swisher, Jacobson, Jun, & Balci, 2001) and health economic models (J. Caro, Ward, & Moller, 2006; J. J. Caro et al., 2006; Ward et al., 2007) (see other applications of simulation in healthcare in Eldabi, Paul, & Young, 2007; Fone et al., 2003; Jun, Jacobson, & Swisher, 1999; Katsaliaki & Mustafee, 2011). Nonetheless, one of the leading barriers to building a simulation model is that it can be a time-consuming process requiring a detailed analysis of the system (Thorwarth & Arisha, 2009).

Kelton et al. (2014) categorised simulation models into three main classes: static vs dynamic, continuous vs discrete and deterministic vs stochastic. In a dynamic simulation, time is continually changing as opposed to the static simulation where time stays constant throughout the simulation. In a continuous model, the state of the system alters continuously such as the amount of water flowing into a tank, while in a discrete model, the state of the system is discrete and changes at a specific point in time (e.g., the number of customers in a bank). Lastly, in contrast to a stochastic model that deals with some level of uncertainty or randomness, there is no uncertainty or random input in a deterministic model. For example, the interarrival time of customers to the bank in the previous example is not deterministic and can be modelled by a probability distribution.

In another classification, there are three types of simulations, namely Monte Carlo, discrete-event and continuous simulation. Monte Carlo simulation offers “an alternative to analytical mathematics for understanding a statistical sampling distribution and evaluating its behaviour in random samples. It does this empirically using random samples from known populations of simulated data to track a statistics behaviour” (Mooney, 1997). In a discrete-event simulation, the state variable changes at a separated point in time (Fishman, 2001). As opposed to discrete-event simulation, in continuous simulation, the state of the
system varies continuously and time is modelled as a continuous flow (Béchard & Côté, 2013). It is also worth mentioning that, both continuous and discrete-event simulations are dynamic models as the state of the system changes over time (Sokolowski & Banks, 2010). Figure 3.5 shows the position of both discrete-event simulation, and Monte Carlo simulation in modelling of a system.

Moreover, simulation can be a terminating or non-terminating (steady-state) simulation. In terminating simulation, the simulation starts at a defined state or time and ends at the period of interest. However, in non-terminating simulation, we are interested in the steady state of the system in which the statistical behaviour of the system will remain stable during the simulation time (Bandyopadhyay & Bhattacharya, 2014; Cassandras & Lafortune, 2010).

It is also worth noting that, although both DES and System Dynamic (SD) are popular simulation approaches in the Operations Research (OR) field (Pidd, 2004), due to several reasons, mainly the types of our research questions, we have chosen DES rather than SD as our simulation approach. First, in DES, individual entities (patients in our model) move through sets of activities and queues in discrete time as opposed to SD where a system is modelled as a series of stocks and flows that are adjusted in pseudo-continuous time.
Second, DES is able to model plenty of complexities in the system while SD is more concerned with dynamic complexity (Lane, 2000) or in other words provides a bigger picture of the system. The complexity involved in DES is the result of different random processes (e.g., patient arrival or service time processes) as well as the endogeneity in the system (Lane, 2000; Morecroft, Robinson, & Words, 2005). Third, it is generally accepted that SD is mainly used to tackle the strategic problems while the focus of DES is more on addressing the tactical/operational problems (Lane, 2000) (note that nurse staffing is an operational decision that is mainly made by a nurse manager/Coordinator in the medical unit rather than a strategic decision that is predominately made by the hospital’s board or CEO). Fourth, in contrast to SD models that generally represent the deterministic behaviour of the system, the DES models are stochastic (Tako & Robinson, 2009). Fifth, although comparing the performance of the system under different scenarios and with both quantitative and qualitative variables is also possible in SD, the DES is more suitable if one desires to obtain quantitative outputs (waiting times in a queue, resource utilisation rates) for the statistical analysis of the system (Brailsford & Hilton, 2001). Finally, according to Forrester (1960), SD models are “learning laboratories” and are not applied to optimisation or point prediction (Brailsford & Hilton, 2001).

With regards to the choice of DES software, several authors have taken a broad variety of indicators into account such as a consistent and user friendly for user interface, database storage capabilities for input data, an interactive debugger for error checking, interaction via mouse, a troubleshooting section in the documentation (Mackulak, Cochran, & Savory, 1994) or syntactic quality, semantic quality, pragmatic quality, test quality and maintainability (R.T. & Hall, 2003). It can also be seen that some works have developed a guideline (Banks, 1991), a checklist (Hlupic, Irani, & Paul, 1999) or a framework (Kitchenham, Pickard, Linkman, & Jones, 2005) to evaluate different tools and features of DES software. In a recent study, Rashidi and Rashidi (2017) considered thirteen indicators including the general features, visual aspects, coding aspects, efficiency, modeling assistance, testability, software compatibility, input/output, experimental features, statistical facilities, user support, finan-
cial and technical features as well as pedigree and applied two ranking methods of Analytic Hierarchy Process (AHP) and Feature Analysis and Weighted Average Sum (FAWAS) to rank the DES software. As a result, it turned out that Arena has got the first rank among 53 DES simulation software. Likewise, reviewing different simulation papers published in scientific journals or conferences, Dias, Vieira, Pereira, and Oliveira concluded that Arena is the most popular DES software among the researchers.

In comparison to other DES tools, Arena also has the following key advantages that outweigh other software (https://studylib.net):

- An entity-based and flowcharting methodology rather than a code-based programming approach that facilitates modelling of a dynamic process.

- Arena’s flowcharting methodology that makes Arena easier to learn, verify, validate and debug.

- Arena Runtime feature that allows analysts to perform what-if simulation analysis using an Arena model built by someone else.

- Import Visio Flowcharts, AutoCAD drawings plus objects, clipart, pictures and video clips.

- ODBC data compatibility that facilitates import and export data from/to any ODBC data file (e.g., Excel, Access, XML, text, Sequential, LOTUS).

- Visual Basic scripting, automation and macro recorder.

- Real Time modelling that allows the model to run in real time or some multiple of real time, and to communicate asynchronously with external devices or external applications.

- SIMAN simulation language engine that makes Arena models run extremely fast and makes it possible for you to model any complex process.

- OptQuest for Arena optimisation software included.
In Paper II, we, therefore, used Arena discrete-event simulation as a solution approach to the proposed queueing model of PART and ICU as we aimed to model the patient flows between ward, PART and ICU. We took our queueing model as an example and used the definitions given by Banks et al. (2010) and Kelton et al. (2014) to explain the main components of a discrete-event simulation as follows:

- **Entities**: are the dynamic object of interest or “players” that move in the system. The entities of our queueing model are patients who arrive into the system, are served by ward, PART or ICU and finally get discharged.

- **Attributes**: are the characteristics of entities such as the ICU priority score or the EWS score that are attached to patients.

- **Variables**: are the characteristics of the whole system that vary during the simulation such as the number of patients in the ICU queue or number of busy nurses in the ICU.

- **Resources**: are any asset, staff, money, material or machine that is seized and then released by the entities. The critical care nurses in our model are counted as resources.

- **Queues**: are waiting lines such as the ward or ICU queue.

- **Events**: are the immediate occurrence that may change the state of the system, variables or attributes. For example, patients’ arrival into or departure from the ICU are examples of an event.

- **Simulation clock**: is a variable that holds the present value of time in the simulation.

We also briefly explain how Arena operates when is executed. Arena stores an event calendar variable that joins both the pending and future event list. In the current list, Arena keeps track of all entities that will return to the system at present. In the future event list, all the events with their scheduled times of completion and type of event are stored. Once Arena is executed, the first event sitting at the top of the event calendar is called (the events
are ordered ascendingly in time). The first event will be removed from the calendar event once it is completed. The next event in the calendar event will be called subsequently, and these steps will be repeated till all the events in the calendar event get executed. During the execution of events, Arena updates the variables, attributes and state of the system.

In regard to the number of repetitions in a discrete-event simulation, we followed the steps proposed by Banks et al. (2010):

1. Run the simulation model for a sample of $R_0$
2. Calculate the sample standard deviation $S_0$
3. Determine the level of error $\epsilon$ and uncertainty $\alpha$ (e.g., $\epsilon = \alpha = 0.05$)
4. The number of iterations required for the simulation is $R \geq \left( \frac{z_{\alpha/2}S_0}{\epsilon} \right)^2$, where $z_{\alpha/2}$ is the $z$-value of the $(1-(\alpha/2))$ percentile of the standard normal distribution.

Lastly, we review the Replication/Deletion Method applied in Paper II according to the approach presented in Alexopoulos and Seong-Hee Kim (2002):

This approach runs $k$ independent replications, each of length $l+n$ observations, and discards the first $l$ observations from each run. One then uses the i.i.d. sample means

$$ Y_i(l, n) = \frac{1}{n} \sum_{j=l+1}^{l+n} X_{ij} $$

to compute the point estimate

$$ \bar{Y}_k(l, n) = \frac{1}{k} \sum_{i=1}^{k} Y_i(l, n) $$

and the approximate $1-\alpha$ confidence interval for $\mu$ having the form

$$ \bar{Y}_k(l, n) \pm t_{k-1,1-\frac{\alpha}{2}} \sqrt{\hat{V}_R(l, n)/k} $$

where $\hat{V}_R(l, n)$ is the sample variance of the $Y_i(l, n)$.
3.6 Regression Analysis

Regression is a statistical technique that explores the effect of one or more variables (called independent variables, predictor variables or regressors) on other variables (called dependent variables or response variables) (Sen & Srivastava, 2011). The term “regression to the mean” was initially used by Francis Galton in 1886 as a result of investigating the relationship between the heights of individuals and the heights of their parents (M. P. Allen, 1997). To analyse a particular pattern in a data, different types of regression models have been developed so far, such as simple linear regression, non-linear or polynomial regression, stepwise or multiple regression, logistic regression (Freedman, 2009), ridge regression and lasso regression models (Tibshirani Robert, 2011).

3.6.1 Linear Regression

In simple linear regression, we study the linear association between a dependent variable and an independent variable, while in multiple linear regression statisticians explore if there is a linear relationship between a response variable and more predictor variables (Freedman, 2009) which is still different from a multivariate linear regression in which multiple correlated response variables are estimated, rather than a single scalar variable (Rencher & Christensen, 2012). We first start with looking into the simple linear regression and in the next section we will discuss a generalisation of ordinary linear regression called Generalised Linear Model (GLM) as we applied Poisson and negative binomial regressions of this family in Paper I.

According to Pardoe (2012), a simple linear regression for a sample of $n$ pairs of $(x, y)$ values, which we denote $(x_i, y_i)$ for $i = 1, \ldots, n$ is expressed as follows:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad (3.5)$$

Where $y$ is a dependent variable and $x$ is an independent variable. $\beta_0$ is called an intercept (the value of $y$ when $x = 0$) and $\beta_1$ is the slope (the change in $Y$ for one unit change in $x$). The error term is also represented by $\epsilon$. 


The main purpose is now to obtain the estimated values $\hat{\beta}_0$ and $\hat{\beta}_1$ for the parameters $\beta_0$ and $\beta_1$ that produce the best fitted line to the sample data. Specifically, we aim to construct the following equation:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\epsilon}_i,$$  \hspace{1cm} (3.6)

Where $\hat{y}$ is the estimated value of $y$ and $\hat{\epsilon}_i$ is the estimated random error in the sample, called a residual ($\hat{\epsilon}_i = y - \hat{y}$). To obtain the values of $\hat{\beta}_0$ and $\hat{\beta}_1$, we minimise the Residual Sum Square (RSS) as follows:

$$\text{RSS} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

We can now derive the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ by solving the above quadratic equation:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Similarly, for a sample of $n$ sets of $(x_1, x_2, \ldots, y)$ values, denoted as $(x_{1i}, x_{2i}, \ldots, y_i)$ for $i = 1, \ldots, n$, a simple linear regression could be extended to a multiple linear regression as below:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \epsilon_i,$$  \hspace{1cm} (3.7)

That can be generally presented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon,$$  \hspace{1cm} (3.8)

Likewise, we should obtain the values of $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$. In other words, our aim in least square regression is to fit a hyper–plane into $(k + 1)$ dimension space that minimises RSS. Instead of writing all the equations, we can present a multiple linear regression as a form of vector and matrix as:
We can now represent the multiple linear regression as a form of matrix:

\[ Y = X\beta + \epsilon \]

The least-squares parameter estimates \( \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k \) are the vectors that minimise the following equation:

\[ \text{RRS} = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = (Y - X\hat{\beta})' (Y - X\hat{\beta}) \]

Solving this equation will result in deriving the least-squares estimator of \( \beta \):

\[ \hat{\beta} = (X'X)^{-1} X'Y \]

Also, the residuals can be presented as:

\[ \hat{\epsilon} = Y - \hat{Y} = Y - X\hat{\beta} \]

It is also worthwhile to illustrate the key assumptions in a linear regression model described by Freedman (2009) and Yan and Su (2009): 1. Linearity: This means the relationship between the dependent and independent variables has to be linear. 2. Homoscedasticity: This assumption highlights that the probability distribution of residuals has constant variance. 3. Independence of errors: This conjectures that the value of \( \epsilon \) for one observation is independent of the value of \( \epsilon \) for another observation. 4. Normality: This presumption tests if the probability distribution of \( \epsilon \) at each x-value is normal and 5. Residuals with mean zero: This assumption checks if the probability distribution of \( \epsilon \) at each x-value has a mean of zero.
3.6.2 Generalised Linear Model

To address some of the limitations of the classical linear regression such as the response variable should be continuous or at least quasi-continuous, [Nelder and Wedderburn (1972)] introduced GLMs. GLMs are developed for regression models with non-normal dependent variables [G. Dunteman & Ho. (2006)]. They have three main components: 1. aside from the linear regression part of the classical models, different distributions from special family, exponential dispersion models (e.g., Gamma or Weibull) can be modelled, 2. they employ transformations of the mean, named a “link function”, that links the regression part to the mean of one of these distributions and 3. a linear predictor [Lindsey (1997)]. For example, ICU or hospital length of stay can be modelled as count variables through Poisson or negative binomial regressions. Or, the ICU mortality is commonly modelled as a binary variable (0 = death, 1 = survive) in the logistic regression. In GLM, the predictor variables can also take a broad range of non-normal distributions such as uniform distribution. Besides, unlike one of the central assumptions of linear regression that residuals should be normally distributed with mean zero, for binary variables and count variables, the variation about the conditional mean is a function of the mean [G. Dunteman & Ho. (2006)].

Firstly, we started with the logistic regression and used the definitions and notations provided by [Hosmer and Lemeshow (2000)]. In the linear regression model, we saw that

$$E(y \mid x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$$

In the logistic regression, we represent this conditional mean as $\pi = E(y \mid x)$ where $\pi$ is an specific form of logistic regression and defined as:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k}}$$

The transformation of $\pi(x)$ is the logit transformation and can be defined as:

$$g(x) = \ln\frac{\pi(x)}{1 - \pi(x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

Unlike the linear regression where $\epsilon$ has a normal distribution with mean zero and constant variance, in the logistic regression the residuals can have two outcomes: if $y = 1$, ...
then $\epsilon = \pi(x) - 1$ with probability $\pi(x)$, and if $y = 0$, then $\epsilon = -\pi(x)$ with probability $1 - \pi(x)$ (G. H. Dunteman & Ho, 2005). Thus, odds can be defined as:

$$\text{odds} = \frac{\pi(x)}{1 - \pi(x)} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k} = e^{\beta_0} e^{\beta_1 x_1} e^{\beta_2 x_2} \ldots e^{\beta_k x_k}$$

Also, odds ratios in the logistics regression are $e^{\beta_1}, e^{\beta_2}, \ldots, e^{\beta_k}$ for $x_1, x_2, \ldots, x_k$ variables.

Secondly, with regards to modelling of HOSLOS in Paper I, we particularly discuss the Poisson regression as the standard model of count data. Finally, we present the negative binomial and log-linear models that could be considered as other alternatives in modelling of length of stay. We follow the definitions and notations of Cameron and Trivedi (2013).

A discrete random variable $Y$ has a Poisson distribution with rate parameter $\mu > 0$ and exposure $t$ (length of time during which the events take place) with following density:

$$P_r[Y = y] = \frac{e^{-\mu t} (\mu t)^y}{y!}, \quad y = 0, 1, 2, \ldots$$

where $E(Y) = \text{Var}(Y) = \mu t$.

We can now define the Poisson regression for $n$ independent observations, the $i^{th}$ of which is $(y_i, x_i)$. The response variable $y_i$, is the number of occurrences of the event, and $x_i$ is the vector of linearly independent predictors. A regression model can be defined through conditioning of the response variable $y_i$ on $k$-dimensional vector of independent variables, $x'_i = [x_{1i}, \ldots, x_{ki}]$ and coefficients $\beta$ such that $E(y_i | x_i) = \mu(x_i, \beta)$. We can now conclude that, $y_i$ has a Poisson distribution with density

$$f(y_i | x_i) = \frac{e^{-\mu_i} (\mu_i)^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots$$

where

$$\mu_i = E(y_i | x_i) = e^{x'_i \beta} = e^{x_{1i} \beta_1 + x_{2i} \beta_2 + \ldots + x_{ki} \beta_k}$$

The negative binomial regression is another model of count data which is analogous to multiple regression, but the response variable $(y_i)$ is distributed according to the negative binomial distribution. Also, although the negative binomial regression is considered as a
generalisation of the Poisson model, its mean and variance is not equal. We now present the results obtained by Cameron and Trivedi (2013) and Hilbe (2011) as follows:

In the Poisson distribution, if we include a gamma noise variable with a mean of one and a scale parameter of \( \nu \), we can then define the negative binomial (Poisson-gamma mixture) distribution with a mean of \( \mu \) and exposure time \( t \) as:

\[
P_r[Y = y_i | \mu_i, \alpha] = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\Gamma(\alpha^{-1})} \left( \frac{\alpha^{-1}}{\alpha^{-1} + \mu_i} \right)^{\alpha^{-1}} \left( \frac{\mu_i}{\alpha^{-1} + \mu_i} \right)^{y_i}
\]

where

\[
\mu_i = t_i \mu \\
\alpha = 1/\nu
\]

Similar to Poisson regression, for \( k \) predictor variables of \( x \), the mean of \( y \) in the negative binomial regression is presented as:

\[
\mu_i = e^{\ln(t_i) + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}}
\]

Usually, \( x_1 \equiv 1 \), so \( \beta_1 \) as an intercept. Replacing now \( \mu_i \) in the above conditional probability, the negative binomial regression can be shown as below:

\[
P_r[Y = y_i | \mu_i, \alpha] = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i + 1)} \left( \frac{1}{1 + \alpha \mu_i} \right)^{\alpha^{-1}} \left( \frac{\alpha \mu_i}{1 + \alpha \mu_i} \right)^{y_i}
\]

The last model that we present here is the log-linear regression model. In general, the logarithmic transformation function is used when there is a non-linear association between the response and predictor variables. As Christensen (1997) argued, the two primary benefits of applying log-linear models are their flexibility in analysing the regression and variance as well as their simple interpretation of odds ratio. The log-transformed of variables with a heavy tail distribution such as ICU LOS (Al Tehewy, El Houssinie, El Ezz, Abdelkhalik, & El Damaty, 2010; Nathanson et al., 2007; Rapoport et al., 2003; Verburg et al., 2014) is another application of log-linear models that makes the analysis of the model and results more convenient.
The log-linear regression model for a response variable $y$ and a predictor $x$ is expressed as follows:

$$\log(y) = \beta_0 + \beta_1 x + \epsilon$$

where the log is a natural logarithm ($e$). To interpret $\beta_1$, we can write the above equation as:

$$y = e^{\beta_0 + \beta_1 x + \epsilon}$$

and take a differentiate to obtain the marginal effect:

$$\frac{dy}{dx} e^{\beta_0 + \beta_1 x + \epsilon} = \beta_1 y$$

$$\beta_1 = \frac{dy}{dx} \frac{1}{y}$$

In other words, one unit increase in $x$, multiplies the expected value of $y$ by $e^{\beta_1}$. It should also be noted that the R-squared reported for the Poisson and negative binomial models are Pseudo R-squared that all attempt to have similar functionality as R-squared in OLS, but in fact none of them is the same as the R-squared in OLS (Nagelkerke $R^2$ or McFadden’s $R^2$, for example, are the two common R-squared in Poisson and negative binomial regression models).

The performance of the model (goodness-of-fit) could also be measured either by the Pearson’s chi-squared or by the deviance (Log-likelihood ratio) tests (note that the Pearson’s chi-squared should be only used for analysing group data). According to G. Dunteman and Ho, (2006), the Pearson’s chi-squared statistic is defined as:

$$P_p = \sum \frac{(y_i - \mu_i)^2}{\mu_i}$$

where the numerator represents the squared difference between observed and fitted values, and the denominator is the variance of the fitted value. In large samples, $P_p$ is approximately distributed according to chi-square with $n-p$ degrees of freedom ($P_p \sim \chi^2_{n-p}$), where $n$ is the number of observations and $p$ is the number of parameters.

The deviance is also defined as:

$$D_p = 2 \sum \left\{ y_i \log \left( \frac{y_i}{\mu_i} \right) - (y_i - \mu_i) \right\}$$
where the first expression is similar to the binomial deviance, indicating “twice the sum of observed times log of observed over fitted”. The second term also represents the sum of differences between observed and fitted values. Likewise, in large samples, $D_p$ nearly follows chi-square with $n-p$ degrees of freedom ($D_p \sim \chi^2_{n-p}$). Thus, when the p-value is less than the significant level ($p < \alpha$), we can conclude that the model does not fit the data (reject $H_0$). In Paper I, we report the Log-likelihood ratio test for the Poisson and negative binomial models (see tables 4.5 and 4.8).
Chapter 4

Paper I

ICU Nurse Shortage: An Econometric Analysis to Support Allocation of Nurses between PART and ICU
Ali Haji Vahabzadeh, Valery Pavlov and Alex Kazemi

Abstract

An Intensive Care Unit (ICU) provides care to life-threatened patients and is equipped with limited and extremely expensive resources. Therefore, it is vital for hospital management to efficiently manage these resources and particularly allocate them to those who benefit most from receiving them. Besides admission and discharge policies, most hospitals have also recently implemented a Patient-At-Risk-Team (PART). The primary goal of PART is to avert ICU admissions by providing critical care to patients in need in a ward. However, PART has to be staffed with ICU-level nurses, but in practice it is often possible to reallocate an ICU nurse to PART. So, the managerial concern is how to split critical care nurses between ICU and PART optimally. On the academic side, this research contributes to the staffing problem in the acute care setting. Our analytical approach also helps managers to derive more meaningful insights from the available data. We use econometric models to estimate the marginal values of a nurse at ICU and at PART, and use those values along with occupancy data to estimate the effect of moving one nurse from ICU to PART and
vice versa. We found that when the ICU nurses are busy the patient hospital length of stay increases by 3% which leads to an increase in the overall patient hospital length of stay by nearly 125 patient-days. We also figured out that the high utilisation rate of the PART nurses expands the patient hospital length of stay by almost 70 patient-days. We evaluated different scenarios of nurse staffing between ICU and PART. We found that it is more beneficial to the hospital to move a nurse from PART to ICU. We figured out that the hospital manager could save approximately $300,000 per annum by allocating a new nurse to PART per shift.

**Keywords**: health care; patient flow; rapid response team; critical care; medical emergency team

### 4.1 Introduction

An intensive care unit (ICU) is a special medical unit in a hospital that provides care to those patients with the highest illness severity. It is staffed by specialised healthcare professionals and offers advanced monitoring and medical therapies. It is also the most expensive unit in a hospital. Typically, about 15% to 40% of the total hospital budget is spent on the ICU (Brilli et al., 2001; Halpern, Pastores, Thaler, & Greenstein, 2007; Reis Miranda & Jegers, 2012) while ICU beds encompass less than 10% of the total hospital beds (Berenson & Assessment, 1984). ICUs tend to be highly utilised, often running at capacity, meaning that resource allocation and possible early patient discharge decisions need to be made for each potential admission.

Previous studies have shown that a proportion of ICU admissions may have been preventable with earlier intervention (McQuillan et al., 1998). It is feasible that measures to allow such early intervention and identification of patients with high illness severity may improve ICU resource allocation and reduce the need for premature discharges (which may exacerbate the problem due to readmission of these patients and subsequent longer length of stay (Kareliusson, De Geer, & Tibblin, 2015; Kramer et al., 2013)).
Most of these potentially preventable ICU admissions appear to come from inpatient wards (Chellel et al., 2002; Cullinane, Findlay, Hargraves, & Lucas, 2005; DoH, 2000; Wilkinson, 1999) rather than the Emergency Department (ED). They can be theoretically dichotomized by underestimation and overestimation of illness severity. Those patients whose illness severity was overestimated may not have required ICU admission (false positives) whereas in the case of underestimation, deterioration may have been preventable by earlier diagnosis and intervention (Franklin & Mathew, 1994; Schein et al., 1990) (false negatives). The latter have been shown to have worse outcomes in terms of prolonged ICU length of stay (LOS) (Caffin, Linton, & Pellegrini, 2007), ICU readmission (Tabanejad Pazokian, & Ebadi, 2014) and increased mortality (McGaughey et al., 2007).

An early measure to allow better classification of illness severity and early intervention was the establishment of Critical Care Outreach Teams (CCOT) (DoH, 2000) composed of advanced nurse practitioners, and/or ICU doctors. More recently there has been the development of the Medical Emergency Team (MET) or Rapid Response Team (RRT) (Franklin & Mathew, 1994), which incorporates an emergency ward response ability. These teams normally incorporate ICU staff. A hybrid approach has been the Patient-At-Risk-Team (PART) in Australia and New Zealand using specialist nurses that can review patients regularly to identify early physiological deterioration, help institute appropriate therapy and also participate in the MET team (Lee et al., 1995; Pirret, Takerei, & Kazula, 2015).

Resource scarcity, however, makes establishment of such teams difficult. In the US, for example, the nurse shortage increased from 6% in 2000 to 12% in 2010, and is projected to reach 29% by 2020 (BHPPr, 2002). In 2012, the New Zealand Nurses Organisation (NZNO) also reported shortages of over 120 registered nurse (RN) positions at Auckland District Health Board (DHB) alone. Scarcity of more specialised ICU nurses is even more concerning (Stechmiller, 2002). For example, 57% of US hospitals reported that the highest level of nursing vacancies were in critical care (Buerhaus, Staiger, & Auerbach, 2000). Likewise, the Faculty of Intensive Care Medicine (FICM) in the UK reports that 96% of acute hospitals operated without an adequate level of nursing staff during day shifts in October 2016 (Silva
Studies have shown that lower nurse-to-patient ratios in ICU are associated not only with higher mortality (Brilli et al., 2001; Tarnow-Mordi, Hau, Warden, & Shearer, 2000) but also increased postoperative complications such as pulmonary and infectious complications, and re-intubation (Amaravadi, Dimick, Pronovost, & Lipsett, 2000). Nurse-to-patient ratios in Intensive Care in Australasia are typically 1:1 or 1:2. Thus, nursing shortages have a significant impact on the ability to admit to ICU as the number of ICU beds available for admissions is confined by these ratios. Resources allocated to the specialised ward-based PART nursing team may decrease the need for ICU admission or length of stay once admitted (due to lower illness severity) therefore decreasing the burden on the resource-limited ICU. However, within a restricted hospital budget, funds allocated to such a team cannot be used to increase the number of ICU nurses (and therefore ICU admission capability). To find an optimal solution to this resource allocation problem the hospital should thereby evaluate the “marginal value” of nursing resources allocated to the ICU versus those allocated to a ward-based team such as the PART.

The specific objective of this study is to find both the value of an ICU admission and the value of a PART visit to a patient. To achieve the objective, we will apply a two-step approach as follows: Firstly, we will use ICU and PART nurse utilisation data to estimate the marginal value of a nurse in either ICU or PART. We will follow the approach of Kc and Terwiesch (2011) and S.-H. Kim et al. (2014), using an econometric model to measure the value of PART and ICU visits to a patient. Secondly, knowing the marginal values obtained and the number of patients experiencing different levels of ICU and PART utilisation, we conduct a ‘what-if’ analysis to estimate the overall impact of moving a nurse from ICU to PART and vice-versa.
4.2 Literature Review

Previous operations management literature on managing ICU capacity mainly looked at areas such as ICU admission [S.-H. Kim et al., 2014; Shmueli et al., 2004, 2003] and discharge policies [Chan et al., 2012; Hosseinifard et al., 2014; Kc & Terwiesch, 2011], ICU bed planning [L. V. Green, 2002; S.-C. Kim, Horowitz, Karl K, & Buckley, 2000; Ridge et al., 1998; Romanin-Jacur & Facchin, 1987], the effect of ICUs on patient mortality [Franklin & Mathew, 1994; Frisho-Lima, Gurman, Schapira, & Porath, 1994; Metcalfe, Sloggett, & McPherson, 1997; Shmueli et al., 2004] and the impact of ICU workload on patient outcomes [Chan et al., 2012, 2016, 2014; Kc & Terwiesch, 2009, 2011]. On the contrary, our study specifically investigates the impacts of PART on ICU capacity.

Although there have been a number of studies of the impact of ICU outreach and MET teams in medical publications, there is a relative paucity in the field of operations management. Kc and Terwiesch (2011) studied the implementation of RRT in hospitals and its implications for hospitals in managing ICU capacity. Another recent paper showed that earlier admission of patients to the ICU could improve patient outcomes (Hu et al., 2015).

A number of medical studies show the benefits of early identification of physiological deterioration in decreasing the rate of cardiac arrest (Bellomo et al., 2003; DeVita et al., 2004), in-hospital mortality (Ball et al., 2003; Bellomo et al., 2003) and LOS related to cardiac arrest (Bellomo et al., 2003). However, other studies do not show that implementation of an outreach or MET service improved patient outcomes (Esmonde et al., 2006; McDonnell et al., 2007; NICE, 2007; Williams & Wheeler, 2009), hospital LOS (Priestley et al., 2004), critical care LOS (Subbe et al., 2003), unplanned ICU admissions (K. Hillman et al., 2005), or readmissions (Bellomo et al., 2004).

In terms of the impact of nursing shortages, studies showed improved patient outcomes when nursing staff is increased in medical units (Aiken et al., 2002; Kane, Shamliyan, Mueller, Duval, & Wilt, 2007; Needleman et al., 2002). Particularly with regards to staffing in the ICU, investigators studying data on patient outcomes and nursing levels from 65 ICUs in the UK, predicted that overall survival would improve by 7% if the nurse numbers
per bed increased by 50% (Duffin, 2014). Similarly, other studies show that having fewer ICU nurses increases the ICU LOS and life-threatening complications (Pronovost et al., 2001), in-hospital and 30-day mortality (Cho & Yun, 2009), and hospital costs (Dimick et al., 2001).

As ICUs are almost entirely utilised (L. V. Green, 2002), delay in admitting acutely ill patients may be associated with increased mortality (Chalfin et al., 2007; Young, Gooder Mc Bride, James, & Fisher, 2003). To create space, the ICU may then choose to expedite its service rate. Kc and Terwiesch (2009) investigated the effect of workload on service time. They showed that workers increased the service rate when the workload increased. Similarly, Chan et al. (2014) examined different scenarios in which this increase in service rate either alleviated or exacerbated the congestion. Concerning resource allocation between the ICU and Step Down Units (SDU), Armony, Chan, and Zhu (2017) concluded that when the cost of abandonment (dying due to the long wait) for critically ill patients is very high, allocating more nurses to the ICU is the optimal policy. However, when this cost is similar to the cost of bumping semi-critical patients from the ICU the optimal policy is to allocate some nurses to the SDU.

Our work contributes to both OM and medical literature. Firstly, unlike Shmueli et al. (2004, 2003), Kc and Terwiesch (2011) and S.-H. Kim et al. (2014), we use the occupancy level of ICU nurses rather than ICU beds as one of the main reasons for the late admission of patients to the ICU or declined admission is the lack of available ICU nurses (Howard, 2005; Litvak et al., 2008a). Secondly, these studies did not examine the impact of PART on ICU capacity (Note that PART can potentially prevent admission of false positives to an ICU which may result in freeing up the capacity of ICU by almost 40% (McQuillan et al., 1998)). Thirdly, our patient severity score, Physiologically Unstable Patient (PUP) score is different from those measured in those papers. The PUP score in our study is composed of the six cardinal vital signs of a patient (respiratory rate, temperature, heart rate, systolic blood pressure, the level of consciousness and urine output) that are measured and ranked between 1 (lowest severity) and 5 (highest severity) for a patient (Figure 4.1). For example,
the severity of illness scores used in S.-H. Kim et al. (2014) are obtained from a lab 24 hours preceding of a patient’s hospitalisation as well as an estimated probability of patient mortality. Lastly, in the medical literature, the MERIT study performed by J. Chen et al. (2015) is best aligned with our research; however, it is important to recognise that our research concurrently incorporates both operational (e.g., the occupancy level of nurses) and patient demographics in the proposed model.

Figure 4.1 PUP Scoring System (Source: PUP scoring system used in Counties Manukau Health in New Zealand)
4.3  The PART-ICU Interaction Process and Data

Description

4.3.1  PART Process

Our data set comes from Middlemore Hospital, one of the largest hospitals in New Zealand, providing treatment to about 100,000 inpatients a year. The hospital has about 800 beds and an ICU unit of 18 beds. The ICU admits approximately 1400 patients per year. The data set contains a 12-month history (1 July 2015 to 30 June 2016) of 8,576 visits of the PART to 2,662 patients across all wards and post-ICU admission outcomes of patients admitted via referral from PART.

All adult inpatients at the hospital have routine measurement of vital signs by ward nursing staff which are then scored for possible physiological deterioration using the PUP scoring system. Frequency of vital sign measurement is dictated by the PUP algorithm (Figure 4.1). Noted in the case of MET call, an ICU resident doctor also accompanies the PART nurses to the ward. The interactions between ward, PART and ICU nurses described in the above are schematically presented in Figure 4.2.

The decision to admit a patient to the ICU is made by the ICU specialist through a conversation that occurs between the MET team members and the ICU specialist. Factors that may affect the decision to admit patients to the ICU include contemporaneous ICU occupancy and relative severity of illness of all patients referred for consideration of admission at that time. Admission to the ICU may not occur if it is felt that such admission would not change the eventual outcome of the patient, in that either illness severity is not high enough that ICU admission is required or ICU therapies are likely to be futile (irreversible disease process or severe chronic co-morbidities) or result in a poor quality of life.

Most of the patients visited by PART nurses require more than one visit. The distribution of number of visits per episode of PART engagement is shown in Figure 4.4. Figure 4.5 shows a density plot of ICU nurse-to-patient ratios from the occupancy data for this period.
4.3.2 Data Description

Data was collected from two hospital databases maintained by the PART and the ICU. These databases are stored confidentially on secure servers. The databases were linked by a unique patient identifier whilst still on hospital servers by researchers working within the hospital. The resulting data was then de-identified by completely removing any data
which could be used to subsequently identify individual patients prior to any analysis. The PART database variables captured include patient demographics (age, sex), hospital admission date/time and discharge date/time, visit date/time, reasons for PART visit, discharge date/time from PART, PUP score at each visit and the decision outcome at the end of each episode.

The patient information in the ICU data set included basic demographics (age, sex), Acute Physiology and Chronic Health Evaluation (APACHE II) score, ICU admission date/time, ICU discharge date/time, ICU outcome (survival or death) and the daily occupancy level of ICU nurses. The APACHE II score is a severity score that is calculated for ICU patients from variables taken from the first 24 hours of ICU admission. It is an integer score from 0 to 71 representing increasing illness severity and it has been well correlated with ICU risk of death (Knaus et al., 1985). It is routinely used for adjustment of mortality ratios for severity of illness.
For each ward patient visited by PART, there was at least one episode ID and one visit number. Each hospital admission could have more than one episode of PART involvement made up of several visits. Every episode was initiated by either direct nurse/doctor referral.
or a call arising from the PUP algorithm and terminated by outcomes such as patient death or transferring to the ICU/HDU. On subsequent visits during the same episode a new PUP score would be assigned and further treatment would depend on this PUP score as well as assessment at that time by the PART.

In terms of the occupancy level of PART nurses in the hospital, there usually are two PART nurses per day (07:00 – 19:00) and night (19:00 – 07:00) shift. Both PART nurses typically visit a patient together if they receive one call at a time for a patient in the ward. However, if they receive more than a call from the ward, they have to visit the patients individually. In situations that the two PART nurses are already in the service and another request is received for PART, the call will likely be served with some delay.

Feasibly, delay in responding to the needs of the acutely ill patient due to competing calls and PART nurse occupancy might impact outcomes. Previous studies have shown that delayed MET calls are associated with an increase in the probability of unplanned ICU admissions and deaths (J. Chen et al., 2015; Trinkle & Flabouris, 2011). Therefore, to examine the outcome of patients who experienced some delays in PART visiting, we counted the number of patients in PART’s service as depicted in Figure 4.6. Although in 40% of the cases, PART had no patient in service, it is evident that in almost 8% of the cases PART had more than two patients in service.

During the period studied, 243 patients were admitted to the ICU through PART and 56 patients were admitted via direct referrals from the ward team (not seen by PART). This latter case arises when the ward team feel that illness severity or trajectory may necessitate

<table>
<thead>
<tr>
<th>Variable</th>
<th>With PART</th>
<th>Without PART</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>55.96</td>
<td>54.18</td>
</tr>
<tr>
<td>Sex (Female)</td>
<td>24%</td>
<td>59%</td>
</tr>
<tr>
<td>APACHE.2.Score</td>
<td>17.37</td>
<td>13.57</td>
</tr>
<tr>
<td>ICULOS (Day)</td>
<td>2.86</td>
<td>2.77</td>
</tr>
<tr>
<td>HOSLOS (Day)</td>
<td>17.56</td>
<td>16.17</td>
</tr>
</tbody>
</table>

Table 4.1 Summary Statistics of Patients With and Without PART Visits
ICU admission and therefore consult directly with the ICU specialist irrespective of the PUP algorithm.

Concerning the difference in the ICULOS between these two groups, we particularly illustrate the role of PART in detecting the false positive and false negative patients in the ward in Figure 4.7. In the previous sections, we argued that the false positive patients are not as acutely ill as the ICU patients and can be monitored in another unit other than the ICU with a lower level of care. Indeed, without any further consequences, they can be discharged earlier from the ICU.

We can observe this in Figure 4.7 which shows that patients without PART visits have a shorter ICULOS. On the contrary, false negative patients whose health deterioration was not detected on time tend to be kept longer in the ICU once they get admitted. The false negative patients can be observed on the far right side of the ICULOS axis in Figure 4.7. The Wilcoxon rank sum test’s p-values for the APACHE II score (p-value = 0.004) and ICULOS (p-value = 0.1) of these two groups of ICU patients are also significant which indicate the impact of PART on ICU patient outcomes. However, given that our findings are based on
limited data, particularly on patient illness severity, the results from such analyses should be treated with caution. Table 4.1 compares the summary statistics of these two groups of patients in average in more detail. On the ICU patients who were visited by PART (243 out of 299 patients), Table 4.2 shows the number of patients in each PART episode outcome.

**Table 4.2 Number of Patients in each PART Episode Outcome**

<table>
<thead>
<tr>
<th>Episode outcome</th>
<th>Number of patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transferred HDU</td>
<td>135</td>
</tr>
<tr>
<td>Transferred ICU</td>
<td>108</td>
</tr>
<tr>
<td>Patient Clinically Improved</td>
<td>2101</td>
</tr>
<tr>
<td>Other</td>
<td>318</td>
</tr>
</tbody>
</table>
4.4 Measuring the Value of ICU and PART Visits to Patient Outcomes

In this section, we examine the values of ICU and PART nurses’ visits compared to patient hospital length of stay (HOSLOS). We first discuss measuring the occupancy levels of nurses in both ICU and PART. We then develop our econometric models to estimate the impacts of ICU and PART visits on the patient HOSLOS. Finally, we perform a ‘what-if’ analysis of the nurse allocations between the ICU and PART.

4.4.1 Evaluation of the Marginal Value of a Nurse

To quantify the value of ICU admission for a patient, there are relatively standard patient outcomes used in both medical and OM literature such as HOSLOS, ICULOS, hospital and ICU re-admission rates, as quantified by Kc and Terwiesch (2011) and S.-H. Kim et al. (2014). In addition to these measures both hospital and ICU mortality rates can be measured (Berenholtz et al. 2002; J. W. Thomas 2004). However, mortality is not necessarily a good measure of hospital safety as it depends more on the nature of the patient’s underlying clinical state and the type of intervention than on the safety of the hospital, and its prevention (as a measure of patient safety) contributes to the failure of hospitals to recognise and appropriately manage patients who are naturally at the end of life (K. M. Hillman, Lilford, & Braithwaite 2014). As our hospital is almost fully utilised during the year, we only estimate HOSLOS as a measure of performance and define it as the time interval between admission to and discharge from the hospital in one continuous admission.

The ICU occupancy at our hospital is currently computed based on the available and required number of nurses per shift (07:00 – 19:00). Our approach takes into account patient acuity which may affect the nurse-to-patient ratio used in the ICU. This approach is more realistic because we have incorporated the severity of every ICU patient in our computations. Similar to Kc and Terwiesch (2011), we also measured the ICU occupancy during
the times of both admission (ICUNurseOCCAdm) and discharge (ICUNurseOCCDis) of
patients to and from the ICU. Our data shows that the nominal means of ICUNurseOCCAdm and ICUNurseOCCDis are 95% and 100%, respectively. Nevertheless, to ascertain whether effective nurse occupancy influences the ICU admission decision, we examined the ICU utilisation under different levels of occupancy (e.g., 75%, 85% and 95% or 80th, 90th and 95th percentiles of the ICU occupancy distribution). We define the ICU as “Busy” (ICUBusyAdm = 1) if the ICU/HDU nurse occupancy during the time of admission of a patient is above the 80th percentile of its occupancy distribution ($\chi^2(1, 3640) = 0.4$, p-value = 0.05). Figure 4.9 demonstrates that at the 80th percentile ICU occupancy impacts the admission of patients with PUP 2 and PUP 5. It is unclear from this retrospective analysis whether patients with the highest illness severity are not admitted during periods of high ICU occupancy due to reasons of possible medical futility or whether there is delay and subsequent admission due to lack of an available bed. It is possible that patients with lower illness severity may be admitted at lower frequency due to the assumption that they can be adequately monitored or treated in the ward setting allowing the smaller number
of remaining ICU beds to be allocated to competing patients with higher illness severity. These findings seem consistent with those of S.-H. Kim et al. (2014) which show that the ICU admits fewer patients during high occupancy levels. Besides, Figure 4.8 illuminates that, when the ICU nurses are busy, not only do fewer patients get admitted to the ICU, but also more patients are bumped out from the ICU (Kolmogorov-Smirnov’s p-value = 0.002, and Wilcoxon’s p-value = 0.00003). Our observation is also consistent with that of Kc and Terwiesch (2011) who found out that the high occupancy level of ICU beds leads to a decrease in the patient ICULOS.

![Figure 4.9 ICU Admission Rate under Low and High Occupancy Level of Nurses](image)

**Figure 4.9 ICU Admission Rate under Low and High Occupancy Level of Nurses**

We denote the PART nurses as “Busy” (PARTNurseBusy = 1) when the number of patients in the service is greater than two (Note that there are only two PART nurses per shift). To precisely measure this effect we excluded some categories of patients such as those who died in the ward, transferred to the ICU/HDU or another hospital, as well as patients who required palliative care only. The data shows that although PART nurses aim to respond to the needs of ward patients as quickly as possible, their occupancy levels impact the patient HOSLOS.
Table 4.3 Description of the Patient Characteristics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Patient age</td>
</tr>
<tr>
<td>Sex</td>
<td>Binary, =1 if Male, 0 if female</td>
</tr>
<tr>
<td>PUPscore</td>
<td>A PUP score of a ward patient visited by PART.</td>
</tr>
<tr>
<td>APACHE 2 Score</td>
<td>Severity score of APACHE II for patients admitted to ICU</td>
</tr>
</tbody>
</table>

4.4.2 Econometric Model for Patient Outcomes

We start with the following model that estimates the impact of the occupancy level of ICU nurses during the time of admission on HOSLOS of a patient i not admitted to the ICU at that point in time:

$$
\log(HOSLOS_i) = \beta_0 + \mathbf{Y}_i + \beta_2 \text{ICUBusyAdm}_i + \beta_3 \text{Reason.for.Visit}_i + \epsilon_i, \quad (4.1)
$$

where $\mathbf{Y}_i$ is a vector of patient-level variables including the Age, Sex and PUP.Score of a patient i in a ward (See Table 4.3). The variable ICUBusyAdm$_i$ is a binary variable that is 1 when the ICU nurse occupancy during the time of admission for patient i (ICUNurseOCCAdm$_i$) is at the 80$^{th}$ percentile of the ICU occupancy distribution and 0 otherwise. Note that, as a common practice in a regression analysis where categorical variables are encoded as dummy variables, we have also encoded $\text{ICUNurseOCCAdm}_i$ as a dummy variable rather than using the numerical ICU occupancy%. This approach is mainly useful as it makes the interpretation and calculation of the odds ratios more convenient. For instance, creating the two zones of Busy and Not-Busy in the ICU in this research would allow us to compare and interpret the impact of ICU occupancy on patient HOSLOS more conveniently. $\text{Reason.for.Visit}_i$ is a categorical variable that contains the reasons that a ward patient i was visited by the PART nurses (e.g., MET call, cardiac arrest or first visit). The vector of coefficients $\mathbf{\beta}$ measures an estimate for the effect of patient-level variables, $\beta_2$ captures the effect of ICU occupancy during the admission time and $\beta_3$ measures the effect of PART reasons for visit on the patient HOSLOS. The error term is also denoted by $\epsilon_i$.  

We also test the impact of the occupancy level of PART nurses on the HOSLOS of patient \( i \) as follows:

\[
\log(\text{HOSLOS}_i) = \beta_0 + \beta Y_i + \beta_2 \text{PARTNurseBusy}_i + \epsilon_i, \quad (4.2)
\]

where PARTNurseBusy\(_i \) is a binary variable that corresponds to the occupancy level of the PART nurses for the patient \( i \) (note that, similar justification mentioned above for using a dummy variable for ICUNurseOCCAdm\(_i \) can also be stated here). Its value is 1 if the number of patients in service greater or equal to 2 and 0 otherwise. The coefficient \( \beta_2 \) captures the effect of PART nurse occupancy on the patient HOSLOS. Similar to Equation 4.1, the error term is also represented by \( \epsilon_i \). The descriptions of operational variables used in Equations 4.1 and 4.2 are presented in Table 4.4.

**Table 4.4 Description of Operational Variables**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOSLOS(_i )</td>
<td>The length of stay of patient ( i ) in the hospital, in days</td>
</tr>
<tr>
<td>ICUNurseOccAdm(_i )</td>
<td>The occupancy level of ICU/HDU nurses during the time of admission for patient ( i ), 0 otherwise</td>
</tr>
<tr>
<td>ICUNurseOccDis(_i )</td>
<td>The occupancy level of ICU/HDU nurses during the time of discharge for patient ( i ), 0 otherwise</td>
</tr>
<tr>
<td>ICUBusyAdm(_i )</td>
<td>Binary, =1 if the ICU/HDU nurses are busy during the time of admission for patient ( i ), 0 otherwise</td>
</tr>
<tr>
<td>PARTNurseBusy(_i )</td>
<td>Binary =1 if the PART nurses are busy to visit patient ( i ), 0 otherwise</td>
</tr>
<tr>
<td>Reason.for.Visit(_i )</td>
<td>The reason for calling PART to visit patient ( i ) in a ward</td>
</tr>
</tbody>
</table>

As noted above, similar to S.-H. Kim et al. (2014), we have used the logarithmic transformation of the HOSLOS in both equations. Overall, there are two primary reasons to employ this approach (Christensen, 1997; Feng et al., 2014): Firstly, to improve the model fit. For example, if the residuals are not normally distributed then taking the logarithm of a skewed variable may improve the fit by altering the scale and making the variable more “normally“ distributed. In fact, using the log transformation is a common practice among researchers to deal with skewed data as it will reduce skewness and make the data a better approximation of the normal distribution.
As shown in Figure 4.10, the HOSLOS in this study is also lognormally distributed with the skewness 3.6. Once we used the log transformation of HOSLOS, the one sample Kolmogorov–Smirnov’s p-value = 0.2 demonstrates that the \( \log(\text{HOSLOS}) \) is normally distributed. Also, the interpretation of the model is more convenient. In other words, using the log-transformation allows the model to be estimated by linear regression using OLS linear regression rather than MLE.

### 4.5 Estimation Results

The results presented in Table 4.5 highlight that the occupancy level of ICU nurses during the time of admission has a significant impact on patient HOSLOS (see the standard errors that indicate how precise the model’s prediction is, as well as the log-likelihood ratio). As discussed by Austin, Rothwell, and Tu (2002) and Verburg et al. (2014), we estimated Equation 4.1 by both methods: **Maximum Likelihood Estimation** (MLE) and **Ordinary Least Square** (OLS). We compared the estimation results obtained in two generalised linear models with Poisson and negative binomial distributions and a linear regression with...
log-transformed length of stay (Table 4.5). According to Hilbe (2011), Maximum Likelihood Estimation (MLE) estimate model parameters by solving the derivative of the model log-likelihood function, named the gradient, when set to zero. The derivative of the gradient with regard to the parameters is called the Hessian matrix, upon which model standard errors are based. As discussed by Lawless (1987), for the negative binomial regression, there are two first-order equations, one for the model’s coefficients and one for the dispersion parameter. Similar to the Poisson model, the Newton-Raphson procedure or the scoring algorithm could be employed to solve these equations. In this paper, we used R statistical software version 3.3.2 to estimate the model parameters. We applied Wald test to evaluate the significance of model parameters and the log-likelihood ratio and Pearson’s chi-squared tests to assess overall significance and residual. We found out that in all three models, the coefficient estimates ($\beta_2$) for the explanatory variable ($\text{ICUBusyAdm} = 1$) are significant. One explanation for this finding might be that when there is high ICU nurse occupancy patients referred for admission may be kept longer in the ward or other intermediate care unit until an ICU nurse becomes available. The data also shows that the average HOSLOS when ICU is busy ($\text{ICUBusyAdm} = 1$) is 17.03 days as compared to 15.46 days when ICU is not busy ($\text{ICUBusyAdm} = 0$). We emphasise that even though there is an empty bed in the ICU, the patient will not be accommodated in the ICU if there is no ICU nurse available.

We now analyse the impact of the occupancy level of PART nurses on patient hospital length of stay in Equation 4.2. We choose the same approach in estimating this effect that we applied in above. As results show in Table 4.8, the coefficient estimate ($\beta_2$) for the variable ($\text{PARTNurseBusy} = 1$) is significant in all models. The findings indicate that when the PART nurses get busy (there are more than two patients in the service) there may be an adverse impact on outcomes of patients visited by PART. In other words, the utilisation rates of the PART nurses leads to an increase in patient hospital length of stay.


Table 4.5 The Impact of Occupancy Level of ICU Nurses during the Admission Time on Patient Hospital Length of Stay (n = 4334)

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Poisson</th>
<th>negative binomial</th>
<th>log-linear</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICUBusyAdm (Yes)</td>
<td>0.033 (0.009)**</td>
<td>0.022 (0.033)*</td>
<td>0.034 (0.019)*</td>
</tr>
<tr>
<td>Age</td>
<td>0.003 (0.0002)**</td>
<td>0.003 (0.001)**</td>
<td>0.003 (0.0004)**</td>
</tr>
<tr>
<td>Sex (M)</td>
<td>0.191 (0.008)**</td>
<td>0.197 (0.027)**</td>
<td>0.075 (0.016)**</td>
</tr>
<tr>
<td>PUP.Score0</td>
<td>0.056 (0.013)**</td>
<td>0.046 (0.046)</td>
<td>0.020 (0.026)</td>
</tr>
<tr>
<td>PUP.Score1</td>
<td>0.101 (0.013)**</td>
<td>0.096 (0.045)**</td>
<td>0.031 (0.026)</td>
</tr>
<tr>
<td>PUP.Score2</td>
<td>0.078 (0.013)**</td>
<td>0.080 (0.046)*</td>
<td>0.063 (0.026)**</td>
</tr>
<tr>
<td>PUP.Score3</td>
<td>-0.031 (0.015)**</td>
<td>-0.030 (0.051)</td>
<td>-0.008 (0.029)</td>
</tr>
<tr>
<td>PUP.Score4</td>
<td>-0.071 (0.018)**</td>
<td>-0.068 (0.061)</td>
<td>-0.021 (0.035)</td>
</tr>
<tr>
<td>Reason.for.Visit (MET call)</td>
<td>0.199 (0.023)**</td>
<td>0.206 (0.086)**</td>
<td>0.032 (0.049)</td>
</tr>
<tr>
<td>Reason.for.Visit (Other)</td>
<td>0.002 (0.056)</td>
<td>-0.038 (0.193)</td>
<td>-0.203 (0.109)*</td>
</tr>
<tr>
<td>Reason.for.Visit (Routine)</td>
<td>0.115 (0.008)**</td>
<td>0.120 (0.028)**</td>
<td>0.097 (0.016)**</td>
</tr>
<tr>
<td>Reason.for.Visit (SET call)</td>
<td>0.220 (0.123)*</td>
<td>0.198 (0.447)</td>
<td>0.110 (0.254)</td>
</tr>
<tr>
<td>Constant</td>
<td>2.401 (0.016)**</td>
<td>2.372 (0.054)**</td>
<td>0.217 (0.031)**</td>
</tr>
<tr>
<td>Log-likelihood (Pr&gt;χ²)</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td></td>
</tr>
</tbody>
</table>

*Note: Standard errors are shown in parentheses.

* p<0.1; ** p<0.05; *** p<0.01

4.5.1 “What-if” Analysis of Possible Nurse Re-allocation Decisions

On the interpretation of the Poisson model in Equation 4.1, the incident rate ratio is exp (0.033) = 1.03. This means patients when ICUBusyAdm = 1 compared to those when ICUBusyAdm = 0, while holding the other variables constant in the model, are expected to have a rate 1.03 times greater for HOSLOS. Of 1824 patients (a subset of patients who were not sent to the ICU/HDU) in our data, there are 246 patients with ICUBusyAdm = 1. Besides, the average HOSLOS when ICUBusyAdm = 1 is 16.83 days. So, it could be calculated that when ICUBusyAdm = 1, the patients HOSLOS increases by 0.03 * 246 * 16.83 = 124.21 patient-days. It should be noted that we are not saying that the HOSLOS increases by 125 days for an individual patient as each patient has different LOS in our subset. This is why we have multiplied the rate obtained for one patient (0.03) when ICU is
Table 4.6 Number of Patients in Different Statuses of ICU Occupancy

<table>
<thead>
<tr>
<th>ICU Occupancy</th>
<th>Number of patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICUBUsyAdm = 1</td>
<td>246</td>
</tr>
<tr>
<td>ICUBUsyAdm = 0</td>
<td>1578</td>
</tr>
</tbody>
</table>

Table 4.7 Number of Patients in Different Statuses of PART Occupancy

<table>
<thead>
<tr>
<th>PART Occupancy</th>
<th>Number of patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>PARTNurseBusy = 1</td>
<td>207</td>
</tr>
<tr>
<td>PARTNurseBusy = 0</td>
<td>1621</td>
</tr>
</tbody>
</table>

busy by the total average of HOSLOS of all patients in the subset. In other words, when ICUBusyAdm = 1, the HOSLOS of those 246 patients increases by 125 days.

In Equation 4.2 also, the incident rate ratio is \( \exp(0.02) = 1.02 \). This means patients when PARTNurseBusy = 1 as opposed to those when PARTNurseBusy = 0, while keeping the other variables fixed in the model, are predicted to have a rate 1.02 times greater for HOSLOS. To calculate the impact as a patient-day, we subset our data that excludes patients died in the ward, transferred to the ICU/HDU or another hospital and also patients with palliative care only. Of 1828 patients in this subset, there are 207 patients with PARTNurseBusy = 1. In addition, the average HOSLOS when PARTNurseBusy = 1 is 16.55 days. So the computation shows that the HOSLOS increases by \( 0.02 \times 207 \times 16.55 = 68.52 \) patient-days. The total number of patients in each status of ICU and PART occupancy are presented in Tables 4.6 and 4.7 respectively.

We are now keen to investigate the changes that occur in the overall patient hospital length of stay once we move one PART nurse to the ICU and conversely. First, we start by moving an ICU nurse to PART. Currently, there are 1593 patients in ICUBUsyAdm = 0 and 231 in ICUBUsyAdm = 1 states. Also, the PART nurses are busy (PARTNurseBusy = 1) in 1592 visits and below their occupancy (PARTNurseBusy = 0) in 707 visits. Once we move one nurse from ICU to PART, the number of patients in ICUBUsyAdm = 1 state will rise to 430 patients (increased by 199 patient-days). Adding one more nurse to PART will
reduce the number of visits that used to be in the \( \text{PARTNurseBusy} = 1 \) state from 818 to 256 visits. The net change in the patient HOSLOS by adding one nurse to PART is 0.3 days. Given 707 visits in the \( \text{PARTNurseBusy} = 1 \), the patient HOSLOS will decrease by 212.1 patient-days. The results finally indicate that moving one nurse from ICU to PART will decrease the overall patient HOSLOS by 13.1 patient-days.

Table 4.8 The Impact of Occupancy Level of PART Nurses on Patient Hospital Length of Stay \((n = 710)\)

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Poisson</th>
<th>negative binomial</th>
<th>log-linear</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{PARTNurseBusy} ) (Yes)</td>
<td>0.021 (0.009)**</td>
<td>0.022 (0.032)*</td>
<td>0.022 (0.018)*</td>
</tr>
<tr>
<td>Age</td>
<td>0.003 (0.0002)***</td>
<td>0.003 (0.001)***</td>
<td>0.003 (0.0004)***</td>
</tr>
<tr>
<td>Sex (M)</td>
<td>0.190 (0.008)***</td>
<td>0.190 (0.027)***</td>
<td>0.073 (0.016)***</td>
</tr>
<tr>
<td>( \text{PUP.Score 0} )</td>
<td>0.048 (0.013)***</td>
<td>0.043 (0.045)</td>
<td>0.032 (0.025)</td>
</tr>
<tr>
<td>( \text{PUP.Score 1} )</td>
<td>0.120 (0.012)***</td>
<td>0.110 (0.044)**</td>
<td>0.062 (0.025)**</td>
</tr>
<tr>
<td>( \text{PUP.Score 2} )</td>
<td>0.095 (0.013)***</td>
<td>0.096 (0.044)**</td>
<td>0.095 (0.025)***</td>
</tr>
<tr>
<td>( \text{PUP.Score 3} )</td>
<td>−0.024 (0.014)*</td>
<td>−0.026 (0.049)</td>
<td>0.013 (0.028)</td>
</tr>
<tr>
<td>( \text{PUP.Score 4} )</td>
<td>−0.071 (0.018)***</td>
<td>−0.073 (0.060)</td>
<td>−0.012 (0.034)</td>
</tr>
<tr>
<td>Constant</td>
<td>2.500 (0.015)***</td>
<td>2.400 (0.052)***</td>
<td>0.250 (0.030)***</td>
</tr>
</tbody>
</table>

Log-likelihood (Pr> \( \chi^2 \)) | <0.001 | <0.001 |

Note: Standard errors are shown in parentheses.

\( p<0.1; \quad **p<0.05; \quad ***p<0.01 \)

We now examine the impact of moving one nurse from PART to the ICU on the overall patient HOSLOS. Out of 1824 patients, the are 231 patients in the \( \text{ICUBusyAdm} = 1 \) state at the moment. By moving one nurse from PART to the ICU, the number of patients in this state will drop to 111, meaning that the overall patient HOSLOS will decline by 120 patient-days. Right now there are 1592 patients in the \( \text{PARTNurseBusy} = 0 \). The net change in the patient HOSLOS by reducing one nurse from PART is also 0.04 days. This relocation, however, could increase the patient HOSLOS as there is now only one PART nurse in the ward who should look after all the calls. The patient HOSLOS will thereby increase by 63.68 patient-days \((1592 \times 0.04 \text{ patient-days})\), thus decreasing the patient HOSLOS by 120
patient-days on the one hand, and increasing the patient HOSLOS by 63.68, on the other hand, will result in an overall reduction in the patient HOSLOS by 56.32 patient-days.

As the comparisons in the above demonstrate, moving a PART nurse to the ICU will reduce the overall patient HOSLOS by 56.32 patient-days as opposed to 13.1 patient-days reductions gained by moving the ICU nurse to PART. In the scenario that we have analysed, it was assumed that an internal nurse was relocated between the two teams. But if the hospital manager aims to keep the same number of nurses in the PART and reduces the 212 patient-days at the same time, then he needs to hire one new PART nurse who works 24 hours a day, 7 days a week (one extra nurse per 12 hours shift). Besides, assuming that if an average cost per patient-day in our hospital is about $2000, the hospital manager could save almost $424,000 by adding a 24/7 nurse to PART. This potential gain is still noticeable if the manager even decides to hire four nurses with a salary of $70,000 each (See the summary of calculations in Table 4.9). It is worth mentioning that similar analyses could be performed for the other two models (negative binomial and log linear) in both Equations 4.1 and 4.2.

One might also be keen to investigate the changes in the overall patient HOSLOS in other nurse allocation scenarios. For example, what if two or more ICU nurses are moved to PART. In these scenarios, similar to what we have performed above, the number of patients in each state of ICU occupancy and PART occupancy has to be re-calculated. In the next step, the costing of the new staffing should be estimated and finally, the best scenario selected. However, given that there are only two PART nurses in the hospital, we would still think that moving more than one PART nurse to the ICU team cannot be an optimal decision. Paying particular attention to the number of patients in the PART service (Figure 4.6), it is immediately evident how the patient HOSLOS would have been changed if all nurses had been assigned to the ICU (no PART in the hospital). Note that, moving more ICU nurses to PART could be also risky and not beneficial to the hospital. We have argued that due to the limited and expensive resources, ICU usually operates in its high occupancy all times. Specifically, it has been highlighted that because of the severity of ICU patients,
it is vital to keep a nurse-to-patient ratio high in ICU. So, although it is possible to take this scenario into account in our what-if analysis, given that in the current hospital, only in 8% of cases, there are more than 2 patients in the PART services, it would not still seem possible that transferring more nurses from ICU to PART result in better staffing. This could be though considered as a limitation of this study.

4.6 Conclusion and Future Research

In this paper, we have addressed a problem of optimal nurse allocation between ICU and PART in one of the largest hospitals in New Zealand. To get an indication of how PART could effectively impact the ICU capacity, we first looked at the flow of two groups of patients admitted to the ICU: patients who were admitted through PART versus those who were admitted directly. We found that patients without PART visits had lower APACHE II scores and relatively discharged sooner from the ICU compared to the latter group. Besides, we realised that the acutely ill patients who were not visited by PART had a comparatively longer ICU length of stay. Our findings revealed that ICU capacity was influenced by the level of PART involvement. Specifically, these false positive and false negative patients could have been managed more efficiently if they had been initially identified and treated by the PART nurses in the ward.

To gain more insights into the impact of occupancy level of nurses in both PART and the ICU on patient outcomes, we further developed econometric models that estimate the relevant effect on patient HOSLOS. We first figured out that when the ICU nurses are busy (above the 80th percentile of their occupancy distribution), patients are expected to have

Table 4.9 Net Changes by Moving Nurses between PART and ICU

<table>
<thead>
<tr>
<th>Medical Team</th>
<th>ICU-&gt;PART</th>
<th>PART-&gt;ICU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient-day</td>
<td>-212.1</td>
<td>63.68</td>
</tr>
<tr>
<td>ICU</td>
<td>199</td>
<td>-120</td>
</tr>
<tr>
<td>Net-change</td>
<td>-13.1</td>
<td>-56.32</td>
</tr>
</tbody>
</table>
a rate 1.03 times greater for HOSLOS. On the data that we used, this is approximately equivalent to an increase in an overall patient HOSLOS by 125 patient-days. Moreover, when the PART nurses are highly utilised (more than two patients in the service), this rate is expected to be 1.02 which leads to a rise in the patient HOSLOS by nearly 70 patient-days.

The increasing concern of nurse shortage, workload and fatigue on the one hand and the necessity of employing highly qualified nurses to monitor the health condition of critically ill patients, on the other hand, make severe challenges for hospital managers on quantifying the value of each critical care nurse to both PART and ICU. We think that the findings of our research have considerable managerial implications with respect to nurse staffing in the ICU. The results suggest that moving a nurse from the PART to the ICU would be the optimal decision as it could reduce the overall patient HOSLOS by roughly 56 patient-days. In addition, bearing in mind that the average cost per inpatient day is roughly $2000 (Ellison 2016), the hospital manager could save almost $424,000 for 212 patient-days by adding a 24/7 nurse to PART. Specifically, assuming the average salary of a qualified nurse is roughly $70,000 per annum, the cost of employing two additional nurses (approx. $140,000) is offset by the savings from reduced length of patient stay ($424000), leading to overall cost savings of approximately $300,000.

Future studies in operations management could investigate how different sizes and compositions of PART would impact the ICU capacity. The PART in our research was a nurse-led team with a size of two. However, some hospitals in the US and Australia implement a medically-led team or even a multidisciplinary team. It could be of interest to find out what combination of critical care outreach team would be beneficial to both patients and hospital. Another possibility for further research is to develop predictive models to estimate the impacts of delays in detection of acutely ill patients by PART on patient outcomes such as ICU re-admission and mortality rates. In this study, we used the PUP scoring system with specific criteria and thresholds. We finally recommend that research is needed to determine how different designs of the early warning scoring system affect the rates of false positive and false negative.
Chapter 5

Paper II

A Discrete Event Simulation Model for Nurse Staffing in Patient-At-Risk-Team and ICU

Ali Haji Vahabzadeh and Valery Pavlov

Abstract

Some critically ill patients in hospital wards may receive sub-optimal care because their deterioration is not identified or not acted upon timely by the ward nurses. At the same time, a low-risk patient might be wrongly identified by a nurse in the ward as an acutely ill patient who needs to be admitted to an intensive care unit (ICU). To detect the early signs of critical health deterioration among patients in the ward, and also avert preventable ICU admissions, hospitals implement a nurse-led team called a Patient-At-Risk-Team (PART). However, today’s global nursing shortage causes a serious and complex challenge for hospital managers on determining nurse staffing levels in hospital medical units. In this paper, we use a queueing model to examine the impact of the nurse staffing level between PART and ICU on ICU patient mortality rate. Using operational flow of patients of a large hospital in New Zealand, we build a discrete-event simulation model. Our results suggest that by re-allocating even one nurse from ICU to PART, the ICU mortality rate can decrease by as much as 35%. The results also illustrate why, despite such a dramatic potential benefit
for patients, a decision to re-allocate even one nurse from ICU to PART can be difficult for managers. The reason is that the utilisation of PART nurses tends to be much lower, by an order of magnitude, than utilisation of ICU nurses. Thus, our study sheds light both on the significant role of the PART in hospitals on improving patient outcomes and ICU performance and also provides some insight into the types of challenges it presents for hospital managers.

**Keywords**: healthcare delivery; capacity allocation; patient flow; simulation; rapid response team (RRT); patient at risk team (PART); medical emergency team (MET)

## 5.1 Introduction

The cost of critical care was predicted at £1.4 billion in the UK in 2009, which would roughly constitute 2% of a total National Health Service (NHS) hospital Trust annual budget (Edbrooke et al., 1999). In particular, the data collected from intensive care units (ICUs) throughout the world reveal that ICU staffing, primarily nursing, accounts for 50% to 80% of direct costs in ICUs (Berenson & Assessment, 1984; Civetta, 1973; Griner & Liptzin, 1971; McCleave, Gilligan, & Worthley, 1977). Despite the high nursing costs, maintaining appropriate levels of nurse staffing in the ICU ensures delivering high-quality care to life-threatened patients. The findings of clinical research demonstrate that the lower level of nurse-to-patient ratios in the ICU is significantly associated with prolonged ICU length of stay (LOS) (Amaravadi et al., 2000; Pronovost et al., 2001) and higher ICU mortality rate (Kane et al., 2007; West et al., 2014). However, the increasing cost of nursing is not the only concern of hospital managers in providing the best possible care to ICU patients. The global nursing shortage poses unprecedented challenges for policy makers and planners related to establishing optimal nursing in the ICU. According to Bureau of Labor Statistics estimations, there will be more than 1 million jobs for Registered Nurses (RNs) in the US (Richards, 2012) by 2022 largely due to a growing elderly population (Hoover, Crystal, Kumar, Sambamoorthi, & Cantor, 2002) and aging nurses (Juraschek, Zhang, Ranganathan,
& Lin, 2012). The scarcity of critical care nurses in ICUs is even more alarming (Buerhaus et al., 2000) as the nurse-to-patient ratios should not go below one nurse to two patients (Bray et al., 2010).

While ICUs are struggling with resource scarcity, there is a considerable number of acute patients in the ward whose health deteriorations were not identified timely by the ward staff. As a consequence, these patients end up with an ICU admission in a life-threatening condition, and stay in the ICU for a prolonged period of time, while, in fact, they could have been averted from ICU if their deterioration had been detected in advance. Clinical studies particularly illustrate that patients who suffer potentially preventable deaths are almost always in the general wards of hospitals (K. M. Hillman et al., 2014), many of them can become as seriously ill as those in ICUs (D. A. Jones et al., 2012), but such deaths are preventable and not common in areas such as ICU (Angus et al., 2004). Another factor contributing to inefficient use of ICU is patients admitted to the ICU as showing signs of critical conditions but, in fact, could have been successfully treated at the ward. According to McQuillan et al. (1998), up to 40% of ICU admissions are preventable.

To overcome these challenges, health care professionals recommended adoption of an innovative model that expands the critical care services beyond the physical boundary of ICU and detects the early signs of patient health deterioration among acutely ill patients in a general ward. The primary goals of such outreach services are: First, to avert ICU admission by identifying patients who are deteriorating; second, to safely discharge from ICU by following up patients discharged to the ward; and, finally, to prevent admissions of patients who are either not in critical condition despite showing signs or can be treated in the ward (DoH, 2000). Similar concepts but under different titles and formats have been developed throughout the world such as a Medical Emergency Team (MET) and a Patient-At-Risk-Team (PART) in Australia and New Zealand (D. Jones et al., 2008; Lee et al., 1995), a Rapid Response Team (RRT) (K. Thomas et al., 2007) in the US and a Critical Care Outreach Team (CCOT) in the UK (DoH, 2000).

However, the results of PART implementations around the world are controversial
Although in each hospital the specific factors can be different largely due to distinct hospital settings (Pedersen, Psirides, & Coombs, 2016), management commitments (S. S. Scott & Elliott, 2009; Wolfe, 2008), limited resources (McDonnell et al., 2007; S. S. Scott & Elliott, 2009) and educational considerations (K. E. Davies, 2011), the primary, most fundamental reason determining the effectiveness of the PART is the qualification of PART staff. For the team to be effective, it must be staffed with ICU-level nurses but, considering their shortage, the only option available for the hospital manager is most often to re-allocate nurses from ICU to PART.

To investigate these challenges, we developed a queueing model that represents the role and functionality of the PART in hospitals as well as its interactions with the general ward and ICU. Based on operational flow data of patients in one of the New Zealand’s leading ICU, we then built a discrete-event simulation model. The results of our simulation study indicate that by removing a critical care nurse from the ICU and allocating her to PART, the ICU mortality rate significantly decreases as much as 35%, this exactly illustrating the preventive role of PART in identifying the acutely ill patients in the ward as well as sorting out the false positive demands for the ICU. To the best of our knowledge, there is no research in the operations management and medical literature that studies, first, the patient flows between the general ward, the PART and the ICU through a queueing model and, second, the nurse staffing problem between PART and ICU.

The remainder of this paper is organised as follows. We first discuss the relevant literature. Section 3 describes our model and the simulation setting. Section 4 presents our results and discussion. Finally, we sum up by discussing our results and suggesting future research.

5.2 Literature Review

Nurse staffing has been an active area of study in management science and operations research. Previous research specifically examined different nurse allocation policies in
hospitals to improve patient outcomes and nurse workloads. de Véricourt and Jennings (2008) looked into the efficacy of nurse-to-patient ratios mandated by California Bill AB 394 and proposed a finite population $M/M/s//n$ queue where the patients have two states of needy or stable. Their findings demonstrate that the size of the medical units impacts the quality of care significantly. The authors further formulated the number of required nurses in medical units through a closed $M/M/s//n$ queueing system (de Véricourt & Jennings, 2011). The results indicate that there is a possible relationship between the frequency of excessive delay and its impact on patient outcomes. To estimate the actual interdependent dynamics of bed occupancy levels and demands of nursing in an emergency department (ED), Yankovic and Green (2011) developed a finite source queueing model with two sets of servers (nurses and beds). Likewise, they examined how the unit size, nursing intensity, occupancy levels and unit length of stay impact the nurse-to-patient ratios and nurse efficiency. Their findings imply that a specific nurse-to-patient ratio in a small medical unit with more acute patients may result in understaffing while the same ratio in a bigger clinical unit with less severe patients may lead to overstaffing.

Particularly on the applications of queueing theory on managing ICU capacity, Armony et al. (2017) studied a queueing model to find the optimal size of the ICU and HDU. Their results reveal that when the cost of abandonment per patient is very high, allocating more nurses to the ICU is an optimal policy, however, when the two costs of abandonment and bumping are very close, it be would be more optimal to allocate some nurses to the HDU and abandon some critical patients from the ICU. They suggest that hospitals utilise HDUs based on patient heterogeneity, required number of nurses in the ICU as opposed to the HDU and comparative cost of lack of access to care for a critical versus semi-critical patient. To investigate the impact of ICU overcrowding on patient bumping, Dobson et al. (2010) use a Markov chain model in that the states represent the residual LOS of each patient. The experimental results reveal that increasing the ICU capacity decreases the percentage of ICU bumped patients as well as ICU re-admissions. An early analysis of an $M/M/\infty$ queueing model of ICU show that the hourly variation in the arrival rate of patients to
the ICU is not likely to remarkably impact the occupancy level of ICU beds (Collings & Stoneman, 1976). L. V. Green (2002) studied a queueing model to predict the accessibility of beds in the ICU and obstetrics units in New York State. She found that approximately 40% of all obstetrics units and 90% of ICUs have no sufficient capacity when beds are required. The results of a queueing model developed by Chan et al. (2016) show that an increase in the ED boarding times leads to longer ICU LOS. An M/H/c/∞ queueing model of the ICU proposed by Griffiths et al. (2006) in that the high variation in the patient LOS was considered as the main tool for managing ICU resources. To predict the ICU LOS, the flow of patients was modelled as a discrete dice Markov process (Seth Kapadia, Chan Sachdeva, Moye, & Jefferson, 2000). The results demonstrate that there is a high correlation between the ICULOS and utilisation of resources.

Capacity planning and patient flow analysis through simulation modelling are the other areas that have drawn the researcher’s attention to the ICU operational problems. McManus et al. (2004) modelled an ICU as an M/M/c/c queue and applied a spreadsheet simulation technique to find out how different ICU utilisation rates impact the ICU rejection rate. Seshaiyah and Thiagaraj (2011) analysed the interactions between the ICU and general units to identify the sufficient beds required in each unit. Similarly, Romanin-Jacur and Facchin (1987) applied a simulation technique to plan the number of beds and nurses required in a semi-intensive care unit. Using the same approach, Ridge et al. (1998) found that there was a non-linear association between numbers of beds, average occupancy level and number of patients that have to be discharged due to ICU bed shortage. Aiming at improving both ICU utilisation and patient outcomes, S.-C. Kim et al. (1999) analysed various ICU admission and discharge policies through a simulation study. To minimise the number of cancelled elective surgeries due to the lack of ICU beds, S.-C. Kim et al. (2000) further simulated different bed-reservation policies. In a similar context, Kolker (2009) also simulated ICU patient flow to establish a quantitative link between the daily load levelling of elective surgeries and ICU diversion. Hagen et al. (2013), via a priority queuing simulation model, found that prioritising ICU patients based on their severity
significantly reduces delays for critical patients but increases the average waiting time for all patients. Their findings also show that aggressive bumping noticeably increases ICU mortality and readmission rates. Using a similar approach, Mathews and Long (2014) examined the impact of ICU and HDU re-sizing on admission wait times, ICU utilisation by acuity type, and average unit occupancy. Their findings illustrate that as the ICU expands into HDU beds, ICU admission wait times for critically ill patients decrease; and, as HDU size decreases, the waiting times for admission of sub-acute patient increase. Concerning the fact that ED patients can be declined from ICU services due to scarcity of resources, via mathematical modelling and simulation, Litvak et al. (2008a) discussed how several hospitals in a region could share their beds for regional ED patients. To study ICU bed occupancy levels, Mallor and Azcárate (2014) incorporated ICU staff decision rules in a combined simulation and optimisation model.

Although many attempts have been made to model patient flows between ED, ward, HDU and ICU through the queueing theory, we are not aware of any queueing system that particularly models patient flow between PART and ICU. More specifically, the discussion on nurse staffing between PART and ICU through the queueing theory and simulation study has not yet been established in the literature. Despite this gap, our queueing model bears a close resemblance to that of Armony et al. (2017). Our simulation and optimisation approach are also more or less identical to those used in Abo-Hamad and Arisha (2013); Ahmed and Alkhamis (2009); Mallor and Azcárate (2014); Zeinali, Mahootchi, and Sepehri (2015). Specifically, Abo-Hamad and Arisha (2013) applied DES and optimisation models to analyse the patient flow in ED to obtain the optimal nursing in this department. Ahmed and Alkhamis (2009) also applied the same approach to determine the optimal number of doctors, lab technicians and nurses required to maximise patient throughput and to reduce patient time in the system subject to budget restrictions. Mallor and Azcárate (2014) developed a DES model to study bed occupancy levels in an ICU that incorporates the management decisions by clinical staff. Finally, Zeinali et al. (2015) used a DES model to improve the patients flow and relieve congestion by changing the number of ED resources.
(i.e., the number of receptionists, nurses, residents, and beds). The proposed model is used to minimise the total average waiting times of patients subject to both budget and capacity constraints.

5.3 Model

5.3.1 Process Description

Middlemore Hospital is one of the largest hospitals in New Zealand with about 800 beds. The hospital admits approximately 100,000 patients per year and manages nearly 350,000 day-patients and outpatient attendances. The facility has an ICU unit of 18 beds, an ED of 27 adult assessment rooms and 24 beds for short-stay patients. Annually, the ICU admits about 1400 patients. Roughly 39% of ICU admission sources are from ED, 34% from the Operating Room (OR) and 23% from the ward. The remaining ICU sources (about 4%) come from other hospitals or ICUs. Both the ICU and ED operate 24 hours a day, 7 days a week. The nurse-to-patient ratios in the ICU largely depend on patient illness severity but normally is 1:1. For example, some patients who are not so critically ill but still need acute care are monitored by a nurse in the HDU who is also looking after other patients. The reverse is also true. Some severely ill patients in the ICU need to be supervised by more than one critical care nurse. The PART at Middlemore Hospital has two nurses per day (07:00 – 19:00) and night (19:00 – 07:00) shift. The PART nurses are as the ICU nurses but they provide clinical expertise for acute/critically ill patients in the ward rather than ICU.

As Figure 5.1 illustrates, the process begins when a patient arrives at the ED or ward (patients are rarely admitted directly to the ICU, so we disregard this patient flow). Once doctors examine the ED patient, they might be treated shortly in the ED and discharged home. If it turns out that they need more observation they will be transferred to other medical units (e.g., ward). However, if the patient is diagnosed as a critically ill patient requiring ICU services, the ED doctor consults with an ICU specialist on the health condition of the patient and requests an ICU bed. A decision to admit the patient is always made
by the ICU doctor taking into account the severity of candidate ED patient as opposed to the severity of all ICU patients. More importantly, the availability of ICU nurses largely influences the ICU doctor’s decision whether to admit or refuse the patient. In fact, the limiting factor is nurses, and although the number of beds is 18, patient admissions are capped by the number of available nurses, which, according to our dataset is around 11 on average.

All adult inpatients arriving at the ward have their vital signs monitored using the Physiologically Unstable Patient (PUP) scoring system. This system is used to produce scores between 1 (less severe) and 5 (high severe) based on routine observations and examinations performed by ward nurses. During each inspection, the six vital signs (temperature, systolic blood pressure, heart rate, respiratory rate, level of consciousness and urine output) of a patient are measured, and the total score is recorded. If a patient gets an overall score of one (PUP 1), then the nurse in charge will be informed. Also, the frequency of measuring the vital signs will be increased to two-hourly or more if required. If the patient gets PUP 2-4, the frequency of examinations will be reduced to a half hour, and subsequently, a referral to the PART will be triggered if the PUP score remain unchanged over this period. For any patient of concern or with a PUP score of 5 or more, the ward nurse calls 888 and asks for a Medical Emergency Team (MET), Surgical Emergency Team (SET) or a cardiac arrest team. If the PART nurses discern that the patient is deteriorating or acutely ill, they will subsequently contact the ICU specialist asking for an ICU bed. However, not all the ward patients admitted to the ICU come through PART assessment. There is a direct flow of patients to the ICU referred by a nurse or a junior doctor while they are examining their patients at the ward. Similar to the above, the ICU specialist is the only reference who could decide whether to admit the patient. It is worth noting that the patient would be declined ICU admission if it was felt that the treatment would not improve the patient condition or have no effect on the ultimate outcome. Another reason to refuse the patient ICU admission or even discharge them right after admission is that they are a “false positive“ patient. It is worthwhile to mention that although some acutely ill patients might die in the ward,
we are only interested in the patients transferred to the ICU either through PART or direct referrals.

Once the patient is admitted to the ICU, they will be monitored and treated by a team of highly experienced and professional doctors and nurses. The patient will be discharged to the ward for more observation when the ICU team perceive that their health condition is stable. The ICU or HDU discharged patient will be followed up by the PART nurses frequently. Nevertheless, a percentage of patients might be readmitted to the ICU due to worsening of their health condition. In this case, the decision and steps to treat these patients are similar to what we described previously. Finally, some critically ill patients might die due to the severity of their health conditions as shown in Figure 5.1.

Figure 5.1 PART-ICU Process Chart
5.3.2 Data Description

The de-identified data that was supplied by Middlemore Hospital covers a 12-month history (1 July 2015 to 30 June 2016) of 8,576 visits of the PART to 2,662 patients across all wards and post-ICU outcomes of patients admitted through the PART. The data was collected from two separate data sets, the PART and ICU. The PART database includes patient-level information (age, sex), hospital admission date/time and discharge date/time, visit date/time, reasons for PART visit, discharge date/time from PART, PUP score at each visit and the decision outcome at the end of each episode (each episode is made up of several visits). The ICU database also contains patient-level information (age, sex), Acute Physiology and Chronic Health Evaluation (APACHE II) score, ICU admission date/time, ICU discharge date/time, ICU outcome (survival or death) and the daily occupancy level of ICU nurses. The two data sets, however, are linked through a unique patient ID assigned to each patient at the time of admission to the hospital.

![Figure 5.2 PART Arrival Distribution](image-url)
Of 8,576 visits performed by the PART on the ward (Figure 5.2), nearly 31% were assigned PUP 3, 4 and 5 (class $r_1$) and the remaining were given PUP 0, 1 and 2 (class $r_2$). Only about 9% of 2,662 patients visited by the PART were admitted to the ICU/HDU and about 79% were treated in the ward. Besides, as presented in Figure 5.3, the majority of ward patients remained less than one day with the PART. Note that the PART LOS is the time difference between the first PART visit date/time and the discharge date/time from PART follow up visit. Concerning ICU outcomes, 86% of patients clinically improved and had been discharged to the ward, 6% died and the rest were transferred to other units/hospitals.

![PART Length of Stay](https://example.com/part LOS.png)

**Figure 5.3 PART Length of Stay**

The false positives and false negatives were also 14% and 17%, respectively. False positive patients are those patients whose medical conditions were incorrectly identified as requiring ICU services as opposed to false negative patients whose medical conditions were either not recognised as critical or not acted upon promptly enough. To calculate both
false positives and false negatives, we looked at the PUP scores given to a patient and the associated decision made by both nurses in the ward and an ICU consultant (e.g., transfer the patient to ICU/HDU or keep the patient in the ward etc.). The percentage of patients who were assigned high PUP scores of 4, 5 or even MET call, but eventually improved clinically in the ward rather than got admitted to the ICU/HDU are counted as false positives. On the contrary, the percentage of patients who were assigned low PUP scores, but further got admitted to the ICU/HDU are considered as false negatives. Note that, to obtain these ratios, we only relied on the PUP scores and the decisions made exclusively based on the health condition of a patient regardless of the impact of the occupancy level of the ICU on the decision outcome. Lastly, of those ICU patients discharged to the ward, 9.5% were later re-admitted to the ICU/HDU.

![Figure 5.4 ICU Arrival Distribution](image)

**Figure 5.4 ICU Arrival Distribution**

Over this 12-month period, 243 patients were admitted to the ICU/HDU through PART, and 56 patients were admitted directly from the ward. Figure 5.4 compares the ICU arrival
distribution between the two groups. Note that we did not divide our data into weekdays and weekends as there is no association between the ICU mortality and ICU admission days ($P(\chi^2 > 3.41) = 0.75$). Figure 5.5 also contrasts the ICU length of stay distributions of the two groups of patients, accordingly. Similar to most papers that show the weighted tail distributions (e.g., lognormal or Weibull) can be best fitted to ICU LOS distribution (for example, see Kc & Terwiesch [2011] S.-H. Kim et al., [2014]), our analysis also demonstrates that the ICU LOS for both groups is lognormally distributed (Table 5.3). Besides, to reflect the reality of the ICU in that some high severity patients might require more observations, we did not treat any ICU patients as outliers but included all of them in our analysis.

5.3.3 Ward-ICU Queueing Model

To gain more insights into the proposed queueing model, we first discuss the general ward-ICU queueing model, ignoring the existence of PART. In this model (Figure 5.6),
patients arrive to the ward at rate $\lambda_g$ and are served with service rate $\mu_1$. Patients might be also transferred to the ward from some other medical units such as ED or OR but in our model we presume that all the patients arrive at the ward with the same arrival rate. We assume that the number of beds in the ward is infinite ($n=\infty$). That is, there are enough beds to accommodate all arriving patients immediately. Thus, since patients do not have to wait there is no abandonment, and since there are always beds available patients never get bumped from the ward. Also, note that we have used the LOS distributions in both ICU and ward to obtain the service rates ($\mu_1$ and $\mu_2$).

The ward patients are discharged from hospital once their health conditions have improved. However, the health conditions of some patients might deteriorate due to the severity (or type) of their illnesses. These seriously ill patients will be at some point sent to the ICU/HDU and served with service rate $\mu_2$. Upon staying in the ICU they are transferred back to the ward where they (most of them) finally recover.

All patients directed to the ICU are considered equally critical (the same priority), in need of immediate admission. The most important implication of this assumption is that when a new patient arrives and there is no nurse available in the ICU then one of the ICU patients is transferred to the ward (early discharge). In our model, the ICU discharge policy is FIFO. The reason is that, considering that all arriving patients are homogeneous, the patient who stayed longer is the one whose health conditions improved most and is safest to discharge.

When the ICU patients are improving and moving towards recovery, they will be discharged from the ICU and either transferred to the HDU or sent back to the ward. However, their health conditions might get worse, and even in some cases could result in patient death. In the proposed model, the patient’s death is independent of their length-of-stay in the ICU as the death might be due to some complicated medical reasons which are beyond the scope of this research. In the preliminary model, we also assume that the service times for all the patients in the ward whether new or those who were sent from the ICU are indifferent. Unlike other research in which the ICU and HDU are considered
as separate medical units, in this study these two units were merged into a single server with a service rate $\mu_2$. In other words, we are not interested in analysing the patient flow between the ICU and HDU (e.g., admission or readmission from the HDU to the ICU).

In addition, patients can be readmitted to the ICU/HDU owing to several reasons. Worsening of the patient illness, inadequate care or prematurely discharged (demand-driven discharge) can be counted as some of the main reasons for ICU re-admission. It is essential to clarify that although a demand-driven discharge influences the ICU service time and its occupancy level, we are not interested in analysing this phenomenon. Finally, the admission and re-admission flows can be iterated until the patient get discharged from these medial units.

### 5.3.4 The PART Queueing Model

In the proposed PART queueing model (Figure 5.7) a single server with a service rate $\mu_3$, and two classes of patients are contributed to the Ward-ICU queueing model. In this model, patients arriving at the ward with rate $\lambda_g$ are served with a service rate $\mu_1$. However, if the ward nurses detect the vital signs of health deterioration through the PUP algorithm, the PART will be subsequently called to the patient’s bed for further examinations. As a result of the PART investigation, the patients might be treated in the ward or sent to the ICU. In our model, highly severe patients with PUP 3, PUP 4 and PUP 5 are considered as class one

![Figure 5.6 Ward-ICU Queueing Model](image-url)
patients with arrival rate $r_1$ and less severe patients with PUP 1 and PUP 2 are denoted as class two patients with arrival rate $r_2$.

![Figure 5.7 PART Queueing Model](image)

Once the health conditions of ICU patients become stable, they will be either dismissed to the HDU or sent back to the ward for further treatment. Similar to the ward-ICU model, some of the ICU patients might die due to worsening of their health conditions. Furthermore, patients in the ward might be revisited by PART when their health conditions are getting worse. Besides, it is possible that those patients who got recently discharged from the ICU show some signs of health deterioration and therefore require to be examined by PART. Both classes of patients might be revisited by PART if urgent help is needed (these revisits are shown as backward directions in the Figure 5.7).

The queueing policies for both PART and ICU-HDU queues are a priority-discipline queue based on the PUP scoring system and patient severity. Besides, the heterogeneity of patients requiring outreach services might influence the service time, but in this model, we assume that all patients are homogeneous and the variation in these types of services needed for the patients has no effect on the outreach service time. In this model, we assume that patients can be refused ICU admission when there are no nurses available in the ICU. We also model false positive and false negative patients by comparing the PUP scores
reported by the ward nurses with the decision is made by the PART nurses after visiting the patient in the ward.

Table 5.1 Model Parameters

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$</td>
<td>Arrival rate of all types of patients in the ward</td>
</tr>
<tr>
<td>$\lambda_{ED}$</td>
<td>Arrival rate of patients to the ICU from emergency department</td>
</tr>
<tr>
<td>$\mu_1$</td>
<td>Service rate of patients in the ward</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>Service rate of patients in the ICU-HDU</td>
</tr>
<tr>
<td>$\mu_3$</td>
<td>Service rate of PART</td>
</tr>
<tr>
<td>$r_1$</td>
<td>Proportion of class one patients visited by the PART</td>
</tr>
<tr>
<td>$r_2$</td>
<td>Proportion of class two patients visited by the PART</td>
</tr>
<tr>
<td>$n$</td>
<td>Number of nurses in the ICU-HDU</td>
</tr>
</tbody>
</table>

Concerning nurse staffing, it is noteworthy to highlight that, unlike California in the US or Victoria in Australia, where they have mandated nurse-to-patient ratios, there is no regulated nurse-to-patient ratio in New Zealand hospitals. The common nurse-to-patient ratios in the ICU (1:1 or 1:2) might not be established permanently in New Zealand hospitals meaning that critical care nurses are flexible to be relocated from ICU to HDU or conversely whenever needed. Similar to Armony et al. (2017) we also considered a fixed number of critical care nurses $n$. These nurses are equally expert in the way that they can work in either ICU or PART.

We also conjecture that during the time of this study the size of the PART was stable and the team members remained unchanged. To detect and diagnose the patient’s problems as accurately and quickly as possible, the PART might need to perform some medical tests such as a blood test or X-ray. In practice, some of these medical tests might be done in parallel or series, which can also influence the PART service time and performance. Identifying the variety and duration of all the required medical tests, however, needs a deep medical knowledge which is beyond the scope of this research. Table 5.1 represents the notations of the model described above.
5.3.5 Simulation Model

5.3.5.1 Building Simulation Model in Arena

In this section, we describe how we have converted a process presented in Figure 5.1 into an executable simulation model shown in Figure 5.13 through the queueing models depicted in Figures 5.6 and 5.7. Starting from the arrival process, there are two arrival streams, ED patients and ward patients with the arrival rates $\lambda_{ED}$ and $\lambda_{w}$, respectively. As these two streams have different arrival patterns, we have created two separate entities in the simulation model in Figure 5.13 (ED-Patients and Ward-Patients). We first start with the flow of ED patients to the ICU. The ED patients could be classified into high-risk or low-risk patients. As the outcome of ED triage, the high-risk ED patients could be potentially admitted to the ICU. However, the admission decision is influenced by the availability of ICU nurses as well as the severity of ED patients in comparison with the current ICU patients. This is why we have a priority-discipline queue for the ICU in Figures 5.6 and 5.7. In Figure 5.13, the ED triaging decision is shown as a Decide module (ED severity) that categorises ED patients into high-risk (HR-ED) and low-risk ED patients. As shown in Figure 5.13, the low-risk patients are discharged home while the high-risk ED patients who require ICU services wait in the ICU queue until an ICU nurse has become available. The availability of ICU nurses is checked by another Decide module in the simulation model (Are there nurses available). An ICU nurse will be seized if there is any available otherwise the patient has to wait until a nurse could be assigned to them. We have modelled this part using the Seize (Seize ICU Nurse), Release (Release ICU Nurse) and Remove (Remove from ICU queue) modules in Arena (note that patients are removed from the ICU queue once they get admitted).

Similar to ED patients, the severity of ward patients are assessed through a Decide module (Ward Severity). It is usual that a percentage of ward patients are required more observations thereby admitted to the ward. However, those with less severity are discharged home. We have used the LOS in the ward to model the service rate ($\mu_1$). In the simulation model, we have created a Process module (Ward) and assigned the number
of resources, service rate distribution etc. (see Table 5.3). Similar to ED patients, the ward patients could be divided into high-risk (Ward-HR) and medium-risk (Ward-MR) patients once their illness severity is realised. The Decide module (Ward-Discharge-Conditions) is also employed to assign different routing percentages to each of these groups (e.g., 3% Ward-HR, 50% Ward-MR and 47% Discharged). In both Figures 5.6 and 5.7, the patients flow from ward to ICU or home is also shown. However, there is a significant difference between the two queueing models in terms of the way that the ward patients are transferred to the ICU. In the Ward-ICU queueing model, the patients are sent directly to the ICU without PART nurses involvement. But, in the PART queueing model, the patient’s vital signs are first assessed by the ward nurses using an EWS and based on the outcome the PART nurses decide to either keep and treat the patient in the ward or sent them to the ICU. In the simulation model (5.13), we have shown the decision made by the PART based on the EWS as two Decide modules one for the high-risk (EWS-HR) and the other one for the medium-risk (EWS-MR) patients. The primary reason to make this division was to model the false negative and false positive patients that we have discussed in-depth in the previous sections. Briefly, the false positive patients are those whose illness severity were overestimated (the ward nurse assigned a high EWS score to the patient but the PART nurses identified them as a low-risk patient). On the contrary, the health condition of the false negative patient is underestimated (the ward nurse assigned a low EWS score but the PART nurses recognised the patient as the high-risk patient). We have modelled this logic in Arena through the Assign and Decide modules. The two classes of patients arriving to the PART were also shown in Figure 5.7 with arrival rates $r_1$ and $r_2$. The patient LOS with PART is also used to obtain the service rate distribution in PART. Besides, note that the PART nurses might be busy with other patients once they receive a call from the ward. Therefore, we have used a Decide module (Is-PART-Busy?) to model the occupancy level of PART nurses. Once the PART assessment is completed, the nominated patients for the ICU services will be sent to the ICU and the remaining ones will be served in the ward. However, similar to the steps that we have explained for ED patients, the PART selected
patients will not be admitted to the ICU unless there is an available ICU nurse. Patients who were admitted to the ICU either though ED or PART will be served in the ICU with a service rate $\mu_2$. Finally, the acute patients in the ICU might get cured and transferred back to the ward and then discharged home or die due to their health conditions. Some ICU discharged patients might be re-admitted to the ICU due to worsening of their health conditions. These returns/re-admission flows are shown in both queueing models and Arena as backward arrows.

5.3.5.2 Verification and Validation

The simulation model is mainly useful when the mathematical model is too complex to be solved by analytical methods (Fishman, 2001). In fact, the only way to solve the mathematical equations is simulation (Sokolowski & Banks, 2010). In our PART queueing model, it is also obvious that there is no closed-form solution due to several reasons such as the feedback loops or non-markovian arrival and service distributions. Therefore, we used simulation and particularly Arena Rockwell software, version 14.5 to run a steady-state simulation (CI = 95%, with at least 10% relative precision by the Replication/Deletion Method (Law & Kelton, 2000)) for the PART queueing model (Figure 5.7).

Table 5.2 Interarrival Time Distribution

<table>
<thead>
<tr>
<th>Pathology type</th>
<th>Interarrival time distributions</th>
<th>p-value of $\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICU with PART</td>
<td>-0.001 + EXPO (1.61)</td>
<td>&gt; 0.75</td>
</tr>
<tr>
<td>ICU without PART</td>
<td>EXPO (6.57)</td>
<td>0.404</td>
</tr>
<tr>
<td>PART</td>
<td>WEIB (0.0369, 0.869)</td>
<td>0.301</td>
</tr>
<tr>
<td>Ward</td>
<td>3.62 * BETA (0.842, 11.1)</td>
<td>0.121</td>
</tr>
<tr>
<td>Emergency Department</td>
<td>EXPO (5)</td>
<td>0.5</td>
</tr>
</tbody>
</table>

According to the obtained data and patient flow between the general ward, PART and ICU, we set the routing probabilities, arrival rates, service rates and the available resources into the simulation model. We applied the Arena input analyser to estimate the statistical distributions of the arrival rates and service rates as shown in Tables 5.2 and
In order to show the relationships of the parameters with the data, as examples, we have plotted the interarrival time distribution as well as the LOS for ICU patients referred through PART in Figures 5.8 and 5.9.

Table 5.3 Length of Stay Distribution

<table>
<thead>
<tr>
<th>Medical unit</th>
<th>Length of stay distributions</th>
<th>p-value of $\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICU with PART</td>
<td>LOGN (2.98, 4.46)</td>
<td>0.13</td>
</tr>
<tr>
<td>ICU without PART</td>
<td>LOGN (2.72, 4.81)</td>
<td>&gt; 0.15</td>
</tr>
<tr>
<td>Ward</td>
<td>3.62 * BETA (0.842, 11.1)</td>
<td>0.212</td>
</tr>
<tr>
<td>PART</td>
<td>4.53 * BETA (0.764, 9.33)</td>
<td>0.101</td>
</tr>
</tbody>
</table>

Figure 5.8 Interarrival Time Distribution for ICU with PART
With regards to the number of replications \( n \), we followed the formula proposed by Kelton et al. (2014). Firstly, we started with an initial number of replications \( n_0 = 5,720 \) hours (one month) replication length and 72 hours (10% of replication length) warm-up period. Secondly, we obtained the average \( (W_q = 1.13 \text{ hr}) \) and the standard deviation of waiting time \( (s = 1.39 \text{ hr}) \) in the ICU queue from initial number \( n_0 \) replications. Lastly, we applied the following formula to get the number of replications:

\[
    n \approx z^2 \left( 1 - \frac{z}{2} \right) \frac{s^2}{h^2}
\]

Where \( z_{(\alpha = 0.05)} = 1.96 \) (z corresponding standard normal critical value) and \( h \), the half-width from the initial number of replications, is 10% of \( W_q \) that is 0.113. So the number of replications \( n \) will be approximately 581.27 which we rounded up to 582. Also, to ensure
that the simulated model reached steady-state, we plotted the **Work In Process (WIP)** for the Ward’s queue. It is fairly obvious that the queue is not building up which means the model reached steady-state.

**Table 5.4 Validation of Simulation Model by Comparing the Simulation and Real Outcomes**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Actual mean (day)</th>
<th>Simulated mean (day)</th>
<th>p-value</th>
<th>Confidence Interval ($\alpha = 0.05$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICU LOS</td>
<td>2.86</td>
<td>2.92</td>
<td>0.32</td>
<td>[2.42, 3.42]</td>
</tr>
<tr>
<td>PART LOS</td>
<td>0.35</td>
<td>0.41</td>
<td>0.21</td>
<td>[0.344, 0.475]</td>
</tr>
</tbody>
</table>

We verified our simulation model through an interview with the ICU specialist. Besides, we used some of the inbuilt features of the Arena simulation software (e.g., resource status (idle, busy), resource capacity (MR) and number busy (NR)) to justify our model. The primary reason for the verification phase was to make sure that the logic of model was correct. Subsequently, we validated the final results of our model presented in 5.13 through “Comparison Testing“ (Balci, 1997). Specifically, we compared the ICU LOS and PART LOS between the actual data and simulated data. The most explicit test of a simulation model’s validity is to check if the output data obtained in the simulation is closely similar to the output data of the actual system (Law & Kelton, 2000). To compare the two outputs for our KPIs, we applied a two-sample t-test. As shown in Table 5.4, there were no significant differences between the simulation outcomes and the actual values obtained from the real system.

![Figure 5.10 Work In Process in Ward Queue](image-url)
5.3.5.3 Nurse Staffing Optimisation Problem

As discussed previously, it is of interest to hospital managers to find out the optimal nurse allocation between PART and ICU such that it minimises ICU mortality. We used OptQuest for Arena to solve this optimisation problem aiming at minimising the ICU mortality rate. We set in the optimisation model that the total number of nurses in both ICU and PART could not exceed 18 nurses (note that there are only 18 beds). OptQuest facilitates an automatic search for optimal solutions within simulation models, and it works as follows: Firstly, when an optimisation runs, a start-over command will be issued once OptQuest starts the simulation. Secondly, the values of decision variables and resource capacities will be changed based on those determined either by OptQuest or the user for the simulation scenario. Thirdly, OptQuest mandates Arena to execute the first replication. After each replication, the value of decision variables used in the objective function or constraint expression will be retrieved from Arena. The sequence of iterations will be kept running until the specified number of simulations set by the user. OptQuest then uses its search algorithm to construct a new set of values and iterates the simulation run process. Eventually, this iteration will be terminated after a specific amount of time or directly by the user (for more details, see [Fu, 2015]).

5.4 Results and Discussion

Figure [5.11] shows the simulated patient outcomes under the alternative scenarios that we considered. We started running the simulation with the original scenario in which all the critical care nurses are allocated to the ICU and there is no nurse in the PART. In this scenario, as is illustrated in the chart, the number of deaths after ICU discharge is quite high. It can also be seen that the ICU nurses are almost fully occupied (Table [5.5]). However, the number of deaths decreased dramatically (by almost 35%) once we allocated one of the ICU nurses to PART (Scenario 1). This exactly emphasises the potential and critical role of the PART nurses in detecting the early signs of health deterioration of high-risk patients in
the ward which further will result in many less deaths in the ICU. Noting, although the ICU nurses are still operating at their high occupancy levels, their times are devoted to those high-risk patients who have been correctly and promptly diagnosed as acutely ill patients.

The number of deaths continues to decline considerably when we add another two ICU nurses to the PART (Scenario 3). However, the trend remains almost steady in the later six scenarios. Paying particular attention to the occupancy level of both PART and ICU nurses in Scenario 3, we realise that the configuration of 3 PART and 15 ICU nurses is optimal. This setting not only decreases the occupancy level of the ICU nurses by 6%, the PART nurses are also in their low occupancy level which provides them with sufficient amounts of time and energy to act upon the seriously ill patients in a timely fashion as well as preventing false positive patients unfairly occupying ICU beds.

![Figure 5.11 Number of Deaths after ICU Discharge](image)

### Table 5.5 Nurse Allocation Scenarios between PART and ICU

<table>
<thead>
<tr>
<th>Nurse utilisation</th>
<th>Original</th>
<th>Scenario 1</th>
<th>Scenario 2</th>
<th>Scenario 3</th>
<th>Scenario 4</th>
<th>Scenario 5</th>
<th>Scenario 6</th>
<th>Scenario 7</th>
<th>Scenario 8</th>
<th>Scenario 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>PART</td>
<td>0</td>
<td>0.282</td>
<td>0.186</td>
<td>0.129</td>
<td>0.092</td>
<td>0.075</td>
<td>0.065</td>
<td>0.054</td>
<td>0.047</td>
<td>0.042</td>
</tr>
<tr>
<td>ICU</td>
<td>0.983</td>
<td>0.963</td>
<td>0.949</td>
<td>0.927</td>
<td>0.934</td>
<td>0.939</td>
<td>0.943</td>
<td>0.987</td>
<td>0.995</td>
<td>1</td>
</tr>
</tbody>
</table>
It is also observed that the ICU admission rate increases in some countries during the winter due to some viral illnesses such as flu (Appleby, 2018; Garfield et al., 2001; Pearson, Reynolds, & Stickley, 2012). We have therefore examined the effect of seasonality on our nursing configuration between PART and ICU. As shown in Figure 5.12, it is also evident from our data that the number of ICU admissions from ED increased during the winter but dropped significantly in November and March. It is particularly important to note that the ED referrals account for the highest ICU admission rates in our data (almost 40%) meaning that small increase or decrease in the ED referrals could considerably impact both PART and ICU occupancies. To realise how the nurse allocation between PART and ICU is sensitive to the changes in the number of ICU admissions from ED, we have changed the ED arrival rates from EXPO (5) to EXPO (4) (25% increase) and EXPO (7) (30% decrease). As illustrated in Table 5.6, while the ED arrival rate to the ICU has increased, the optimal nurse configuration is still three nurses in PART and 15 nurses in ICU. It is especially interesting to see that the ICU mortality did not decrease even with two nurses in the PART. However, the number of ICU deaths dropped by almost 27% once we added another nurse to the PART. On the contrary, when the number of ICU admissions from ED decreased, the

![Figure 5.12 ICU Monthly Admission from Different Departments](image-url)
optimal nurse allocation scenario would change to one nurse in the PART and 17 nurses in the ICU. The similar analysis could be extended to other seasons, as well. It is worth mentioning that our sensitivity/seasonal analysis would still suggest establishing a PART in a hospital even with one nurse.

We can also think of other scenarios to examine how sensitive the nurse allocation policy is to changes the system parameters such as the service rates of both PART and ICU nurses or the arrival rates of high-risk and false negative patients to the ICU that affect the ICU nurse occupancy. For instance, one could consider a non-homogenous (time-dependent) arrival rates for patients with different illness severities. Similarly a state-dependent service rate or a workload-dependent service rate could replace the current service rates in the simulation model. For example, in Table 5.6 we have already observed how changes in the ED arrival rate to the ICU could potentially impact the nurse occupancy which in turn affect the death rate in the ICU. Therefore, it is crucial to provide reasonably accurate estimates of these parameters.

Table 5.6 Number of ICU Deaths in Different Nurse Allocation Policies between PART and ICU

<table>
<thead>
<tr>
<th>ED arrival rate</th>
<th>Number of PART nurses</th>
<th>Number of ICU nurses</th>
<th>PART nurse OCC</th>
<th>ICU nurse OCC</th>
<th>Number of ICU deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXPO (4)</td>
<td>0</td>
<td>18</td>
<td>0</td>
<td>0.867</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>17</td>
<td>0.021</td>
<td>0.885</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>16</td>
<td>0.012</td>
<td>0.89</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>15</td>
<td>0.06</td>
<td>0.896</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>14</td>
<td>0.05</td>
<td>0.903</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>13</td>
<td>0.03</td>
<td>0.91</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>12</td>
<td>0.002</td>
<td>0.914</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>11</td>
<td>0.003</td>
<td>0.918</td>
<td>11</td>
</tr>
<tr>
<td>EXPO (7)</td>
<td>0</td>
<td>18</td>
<td>0</td>
<td>0.839</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>17</td>
<td>0.025</td>
<td>0.829</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>16</td>
<td>0.009</td>
<td>0.842</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>15</td>
<td>0.016</td>
<td>0.857</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>14</td>
<td>0.006</td>
<td>0.0867</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>13</td>
<td>0.004</td>
<td>0.877</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>12</td>
<td>0.004</td>
<td>0.887</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>11</td>
<td>0.004</td>
<td>0.906</td>
<td>5</td>
</tr>
</tbody>
</table>
Results and Discussion

Another opportunity could be replacing the current ICU admission policy with those discussed by Shmueli et al. (2003) to find out how the ICU abandon rate, as well as ICU mortality rate, would be altered. Also, the impact of different discharge policies such as shorter LOS or greedy policy proposed by Chan et al. (2012) on ICU patient outcomes could be taken into consideration. Note that the changes in the parameters could indeed illustrate to what extent the system performance is robust. In other words, although we have realised that the system performance is very sensitive to changes in ED arrival rates, one might find that the ICU mortality rate is not sensitive to miss-specifications of ED arrival rate at all but the service rate in the ICU. Lastly, the simulation model designed in Arena is presented in Figure 5.13.
Figure 5.13 Arena Simulation Model
5.5 Conclusion

In spite of a common consensus among the medical community on the necessity of constructing outreach services in hospitals, it appears that research has paid less attention to optimal staffing of critical care in hospitals. In this paper, we have filled the gap and examined the impact of nurse staffing policies between PART and ICU on the ICU mortality rate. We have focused on patient flow between the general ward, ED and ICU-HDU medical units. Notably, we have tracked the streams of false positive and false negative ward patients into the PART as these two classes of patients would mainly impact the ICU capacity and mortality rate. This is the first study in operations management literature that explores patient flow between PART and ICU. Simulating the operational flow of patients in one of the largest hospitals in New Zealand, we have found that the PART nurses play a significant role in alleviating the ICU mortality rate. Our findings were aligned with the general goals of the PART as averting preventable ICU admissions to accommodate more capacities for the right acutely ill patients. Precisely, we were able to find that the ICU mortality rate would decrease if ICUs could implement an effective preventive approach along with a corrective one.

We acknowledge that our research has some limitations, which in turn open avenues for future research. First, our study is limited to one hospital with specific settings. Throughout the world, there are various settings of outreach services such as a Rapid Response Team (RRT) in the US or a Medical Emergency Team (MET) in Australia. Some outreach services are a medically-led team, and some are a combination of nurses and physicians. Our queueing model can be applied in those settings to find out how different compositions of outreach services impact the ICU mortality rate. We have only considered the ICU mortality rate, however, other metrics such as the ICU re-admission rate and ICU LOS, which impact the cost and capacity of the ICU, can be taken into account too.

Second, Kc and Terwiesch (2011) and S.-H. Kim et al. (2014) analysed the impact of the occupancy level of ICUs on the ICU length-of-stay under different scenarios. It would be interesting to go through the same approaches and figure out how the PART impacts the
ICU LOS, hospital LOS and patient outcomes. According to [J. Chen et al. (2015)], with every 100 delayed MET calls, we can avoid 13 deaths. Third, using our approach, further studies need to be performed to find out if “delayed PART” response is associated with higher mortality. Last but not least, future research can examine patient outcomes under different levels of PART performance.
Chapter 6

Discussion and Conclusion

6.1 Discussion

The lack of critical care nurses is so vital that it can lead to delay or even refusal of admission of acutely ill patients to the ICU. Hiring new critical care nurses is not always a feasible solution due to the high cost and, more importantly, scarcity of nurses. However, reducing the unnecessary demands for the ICU could be an alternative solution to the restricted capacity of ICU nurses. The medical society, therefore, proposed that establishing a nurse-led PART might be able to reduce the unplanned or preventable admissions from ward to the ICU thereby freeing up some limited and expensive capacities of ICU nurses for patients with life-threatening conditions. Nevertheless, to the best of our knowledge, it is not well studied what nurse allocation policies between PART and ICU would result in best patient outcomes as well as ICU performance. In this thesis, we addressed this question through two distinct but related solution approaches in one of the largest hospitals in New Zealand.

In Paper I, we proposed econometric models to predict the impact of critical care nurse occupancy in both PART and ICU teams on patient HOSLOS. The reason for measuring the HOSLOS is that, in both medical and OM studies the HOSLOS is often measured as a key indicator of hospital performance, this is mainly because the HOSLOS is highly correlated
Discussion and Conclusion

with hospital costs (Cots et al., 2003; Polverejan et al., 2003); therefore hospitals attempt to shorten HOSLOS while keeping efficiency (Siskou et al., 2009). In this research paper, we were specifically keen to estimate the marginal values of a nurse at the ICU and at the PART, and apply those values along with nurse occupancy data to determine the impact of nurse displacement between PART and ICU. Our findings suggest that nurse occupancy in both PART and ICU has a significant impact on patient HOSLOS, but the effect of ICU nurse occupancy on extending the patient HOSLOS is more substantial. Notably, we have found that the patient HOSLOS would increase by approximately 125 patient-days when ICU nurses were busy as opposed to a 70 patient-days increase in HOSLOS when PART nurses were highly utilised.

We have also examined the impact of ICU nurse occupancy on the ICU admission rate. Our results revealed that at the 80th percentile, ICU occupancy impacts the admission of patients with PUP 2 and PUP 5. According to our ICU consultant, the difference in admission of PUP 5 patient is interesting and theoretically should be presented at all levels of occupancy above 80th. Specifically, he thought that the ICU occupancy could feasibly alter outcome for PUP 4 or 5 patients who were not admitted to the ICU. Nevertheless, we have only observed this impact on PUP 2 and PUP 5. The ICU consultant also commented that it is not clear from this retrospective analysis if patients with the highest illness severity are not admitted during periods of high ICU occupancy due to reasons of possible medical futility or whether there is delay and subsequent admission due to lack of an available nurse. This argument could also be valid as we have not had any variables in our dataset that measured the delay in ICU admission and its subsequent impact on patient outcomes. Also the ICU specialist noted that it is plausible that patients with lower illness severity may be admitted at a lower frequency due to the assumption that they can be adequately monitored or treated in the ward setting allowing the smaller number of remaining ICU beds to be allocated to competing patients with higher illness severity. Our findings also confirm the results obtained by S.-H. Kim et al. (2014) that ICUs admit fewer patients during high utilisation rates. Besides, we have found that when the ICU nurses are busy,
not only fewer patients get admitted to the ICU, but also more patients are bumped out from the ICU (Kolmogorov-Smirnov’s p-value = 0.002, and Wilcoxon’s p-value = 0.00003). These results also corroborate those of Kc and Terwiesch (2011) who have figured out that patients have shorter length of stay when the ICU operates in its high occupancy level.

Although we were expecting to observe the impact of nurse occupancy on the ICU admission rate for patients with PUP 3 and PUP 4 too, the results were only significant for PUP 2 and PUP 5. One possible explanation for this observation could be due to the lack of differentiation between ICU patients and HDU patients in the data. The data reflects the physical setting of ICU and HDU at the Middlemore Hospital that considers both ICU and HDU as a single unit (see Figure 1.3). As we discussed in Section 1.3, HDU patients are less critically ill patients with PUP scores usually ranging from 1 to 4 as compared to the ICU patients with a PUP score 5 and more (e.g., MET call or cardiac arrest). Therefore, it could be possible to say that the HDU, similar to the ICU, refuses admission of less severe patients (PUP 2) when it gets busier. In the meantime, these declined patients would be temporarily managed in another medical unit until an HDU bed becomes free. However, the HDU patients sitting at the upper bound of illness severity (PUP 3 and PUP 4) might not be affected by the occupancy as the manager might decide to discharge the less severe HDU or even ICU patients earlier (demand-driven discharge) to accommodate the higher severity HDU patients. Note that as a common practice among ICUs, when the ICU is crowded, to admit a new patient with a higher severity score than the severities of the ICU patients, a less severe patient is normally discharged earlier so that the arriving patient could be assigned a bed in the ICU.

In Paper II, we have developed queueing models to examine the impact of different nurse allocation policies between PART and ICU on ICU patient mortality rate. We have further performed a discrete-event simulation study using the operational patient flows between ward, ED, PART and ICU to gain a better insight into the impact of different nurse staffing between PART and ICU on ICU patient outcomes. The most striking results emerging from our simulation study revealed that the lack of a PART in a hospital might
increase the ICU patient mortality rate. Our findings substantiate previous findings in the medical literature that found implementing any structure of the outreach services in hospitals such as MET, or RRT has led to improving the patient outcomes (Bellomo et al., 2003; M. D. Buist et al., 2002; Konrad et al., 2010; Sharek et al., 2007). However, note that none of these studies has applied an analytical solution or a simulation study and mainly relied on either a retrospective study or a before-after study.

In Section 3.2, we have extensively discussed studies used the queueing theory to model the nursing in the medical units. For example, papers published by Armony et al. (2017); de Véricourt and Jennings (2008, 2011) and Yankovic and Green (2011) have taken different variables including nurse workload into account to obtain the optimal nurse-to-patient ratio. Although, we are not aware of any research in the OM literature that specifically developed a queueing model to analyse nurse staffing between PART and ICU, the model proposed by Armony et al. (2017) for HDU and ICU could be considered as the closest one to our study. In order to decide on the size of the ICU and HDU, they studied the dynamics of patient flows between the ICU and HDU. They proposed a queueing model in which the patients can be abandoned or bumped when there is no space in the ICU. Computing the costs of abandonment and bumping, they concluded that when the cost of abandonment per patient is very high, allocating more nurses to the ICU is the optimal policy and, on the other hand, when these two costs are very close, the optimal policy is to allocate some nurses to the HDU and abandon some critical patients from the ICU. One of the main differences between our model’s assumptions and their assumptions is that we have modelled false positive and false negative patients. False positive patients are those patients whose illness severities are overestimated (too healthy to benefit ICU care) as opposed to false negative patients whose health conditions are underestimated (late realisation of health deterioration). These two classes of patients can largely impact the ICU capacity. Indeed, up to 40% of ICU admissions are preventable (McQuillan et al., 1998). Apart from the potential improvement that can be gained in the capacity of the ICU, patient outcomes can be significantly improved as well. On the importance of the false
negative patients, in particular, it was found that they normally have worse outcomes in terms of prolonged ICU LOS (Caffin et al., 2007), ICU re-admission (Tabanejad et al., 2014) and increased mortality (McGaughey et al., 2007). Therefore, identifying both false positive and false native ward patients by PART nurses could help the ICU to manage its limited and costly capacity more efficiently. Nonetheless, one of the advantages of their model is that they have concurrently considered both readmission and mortality risks as the cost parameters in their optimisation model. On the main similarities between the two models, we have also assumed that patients are abandoned when there is no nurse in the ICU.

Similar to Duraiswamy et al. (1981); Griffiths et al. (2006); Hashimoto et al. (1987) and Mullinax and Lawley (2002), we have also built a simulation model to study the nursing challenge in the ICU. We believed that applying a simulation study in a highly complex and dynamic system such as healthcare where a wide range of variables and entities are involved could provide managers with a valuable intuition to compare the results of different policies in a computer-based setting rather than a real-world environment. Studies have also proved that the applications of DES in healthcare have increased remarkably (Fone et al., 2003; Jacobson, Hall, & Swisher, 2006; Katsaliaki & Mustafee, 2011). It is also significantly important to highlight the fact that due to several considerations and limitations such as patient safety it is not possible to run an experiment in the hospital to find out what nurse staffing between PART and ICU would result in best patient outcomes.

Concerning the settings in the simulation model such as the interarrival times and service times distributions, we had some similarities and differences with previous studies. Although most papers modelled the ICU arrival process as a Poisson distribution (for example, see L. V. Green, 2002; Griffiths et al., 2005; Kolker, 2009; Litvak et al., 2008a) some papers modelled the arriving patterns to the ICU as a non-stationary Poisson distribution (for example, see Dobson et al., 2010; Griffiths et al., 2005; Masterson et al., 2004) due to the changes in the ICU arrival process within the day. We have followed the first approach and found that the ICU arrival process follows the Poisson distribution as the $p$-values of $\chi^2$ test for the ICU arrival process with and without PART were both statistically significant.
Discussion and Conclusion

(p > 0.75 and p = 0.404, respectively). The results of our data analysis also showed that, similar to Chan et al. (2014); Masterson et al. (2004) and Shahani et al. (2008), the ICU LOS in our model is lognormally distributed for both ICU with PART (p = 0.13) and without PART (p = 0.13). With respect to the heterogeneity of ICU arrival patients, we categorised patients based on their referral source to the ICU (e.g., ED and Ward) as well as their illness severities. In the literature both approaches were applied. For example, Adeyemi et al. (2010); Asaduzzaman et al. (2010); Barado et al. (2012) and Cochran and Bharti (2006) categorised the ICU patients based on their referral departments while Armony et al. (2017); Chan et al. (2012); J. Li et al. (2015) and Wharton (1996) grouped them according to their illness severities.

After the verification of the simulation model by the ICU consultant and also benefiting from some of the inbuilt features of the Arena simulation software, we followed the steps proposed by Altiok and Melamed (2001) to validate our simulation model. We have designed statistical hypothesis testing applying Student’s t-distribution to compare the results of our simulation study with the actual data supplied by the hospital. As Law (2006) stated, the most explicit test of a simulation model’s validity is to check if the output data resulting from the simulation is closely analogous to the output data of the real system. For this purpose, we have chosen the LOS in both PART and ICU as the two most important KPIs. The p-values of t-test demonstrated that there were no significant differences between the simulation outcomes and the actual data obtained from the hospital (p_{ICULOS} = 0.32, p_{PARTLOS} = 0.21).

Lastly, it is worth mentioning that there are always ethical concerns associated with any types of medical research. For example, it is essential to highlight that if a study involves any intervention administered to the participant, such as drugs or medicine, or if it involves potentially hazardous substances or use of human blood, body fluids, or tissue samples or if it is considered a clinical trial. However, our study did not fall into any of these categories. In this study, we have been only supplied de-identified data at both patient-level and hospital-level for the purpose of data analysis and modelling of the research questions.
Also, as we have not considered any variables related to ethnicity, our research did not raise any cultural issues. Besides, the conducted study did not involve a collection of information about illegal behaviour(s) which could place the researcher or participants at risk of criminal or civil liability.

### 6.2 Summary of Research Results

Paper I proposed econometric models to estimate the impact of the occupancy level of critical care nurses in both PART and ICU on patient hospital length of stay. Specifically, this paper developed econometric models to predict the marginal values of a nurse at ICU and at PART, and used those values along with the nurse occupancy data, patient illness severities and demographics data to estimate the effect of moving one nurse from ICU to PART and conversely. This paper fills a gap in the body of OM literature for investigating the effects of constructing a nurse-led PART in a hospital on the limited and costly capacity of ICU. Contrary to both OM and medical research, the current paper focused on modelling of nurse occupancy rather than bed occupancy, as the ICU admission decision is mostly confined by the availability of ICU nurses rather than ICU beds. Using both operational and patient-level data from one of the largest hospitals in New Zealand, our models suggested that the hospital could gain more by moving a critical care nurse from PART to ICU. The findings mainly revealed that the hospital could even save approximately $300,000 annually by allocating a new nurse to the PART per shift.

To gain more insights into the impact of critical care nurse occupancy on ICU capacity, we developed a queueing model of PART and ICU in Paper II. We further built a discrete-event simulation model employing real data from the hospital to examine how different nurse allocation policies between PART and ICU impacted ICU patient mortality rate. One of the distinguishing features of the proposed model is that we demonstrated the role of the PART in expanding the capacity of the ICU through identification of both false positive and false negative patients. As far as we know, this is the first study in OM literature...
to analyse patient flow between PART and ICU through an analytical approach. Using the operational data flows between ward, ED, PART and ICU, we have found that PART nurses play a significant role in alleviating the ICU mortality rate. Crucially, the results of our simulation suggested that constructing a PART in a hospital by reallocating even one nurse from ICU to PART can decrease the ICU mortality rate by roughly 35%. Our findings concur well with those streams of medical research that emphasise the role of PART, MET, RRT or CCOT in reducing the mortality rate. Finally, the results of our “what-if” analysis showed that the configuration of 3 PART and 15 ICU nurses is an optimal nurse allocation policy for an ICU unit of 18 beds.

The evidence from these studies implies that both patients and hospitals could benefit from establishing a nurse-led PART as it might result in improving patient outcomes as well as ICU capacity. For example, in the hospital that we studied, there are two nurses in the PART and both the ICU and HDU are equipped with 18 beds. There are no constant nurse-to-patient ratios in the ICU, these ratios being influenced mainly by the availability of nurses as well as the patients’ illness severities. Our findings suggest that the hospital might reduce the ICU mortality rate by allocating a critical care nurse from ICU to PART. Indeed, in an ideal world where there is no nurse shortage and the hospital has sufficient funds to hire more critical care nurses, constructing a PART with three nurses and an ICU with 15 would be optimal. However, realistically, with hospitals’ challenge with nurse scarcity it is almost financially infeasible for them to hire as many nurses as they require. Therefore, the results suggest that taking both resources and budget constraints into account, the hospital can still improve both patient outcomes and ICU capacity by staffing the PART with only one nurse and transferring another PART nurse to the ICU. Finally, given that our findings are based on a limited number of patients (a 12-month history of 8,576 visits of PART nurses to 2,662 patients) from one hospital, the generalisation of our results should be interpreted with caution.
6.3 Contribution to Research

The overarching goal of this thesis was to demonstrate the importance of critical care nursing between PART and ICU on both acutely ill patient outcomes and ICU performance. Previous studies showed that establishing a PART at hospitals could potentially improve ICU capacity through averting unplanned admissions. Nevertheless, the primary challenge for hospitals is that both units require nurses with the same expertise and, considering the nurse scarcity and the fact that hospitals usually operate under very tight budgets, hiring one or two more critical care nurses is infeasible. Indeed, constructing a PART could be only possible by reallocating one from ICU to PART, but such transition immediately reduces the ICU capacity because ICUs are typically required to maintain the nurse-to-patient ratio 1:1 or even 2:1 when the patient’s condition is more critical. To find an optimal solution to this resource allocation problem the hospital should evaluate the trade-off. Thus, the authors aim to provide a series of arguments on how the current thesis contributes to a significantly less explored realm in both medical and OM literature.

In OM literature, the studies performed by Kc and Terwiesch (2011) and S.-H. Kim et al. (2014) are closest to our research. Overall, they investigate the impacts of ICU admission and discharge policies on patient outcomes, ICU LOS and HOSLOS. Although their findings are remarkable, our research is different from them in the following aspects: First, none of these studies modelled the impacts of any types of outreach services (e.g., RRT, MET or PART) on ICU patient outcomes and ICU capacity. As we have elaborated in previous sections, a considerable number of ICU admissions (about 40%) is preventable if and only if they are identified in advance. This is the principal goal of a PART. Therefore, we believe that including the key role of the PART in enhancing the constrained capacity of the ICU in the model reflects more of the reality of current settings in hospitals.

Second, both studies examined the impact of the occupancy level of ICU beds as opposed to our research that contemplates the occupancy level of ICU nurses. We emphasised the fact that the primary reasons for the late admission of patients to the ICU or their refusal is ICU nurse scarcity rather than ICU bed shortage (Howard, 2005; Litvak, van Rijssbergen, ...
We have examined how different nurse utilisation rates impact patient outcomes and HOSLOS. Note that the nurse-to-patient ratios in our data were not fixed but varied based on the illness severity of patients.

Third, we have used the Physiologically Unstable Patient (PUP) score and APACHE II score that measure the illness severity of both ward and ICU patients, respectively. The PUP score is composed of the six cardinal vital signs of a patient (respiratory rate, temperature, heart rate, systolic blood pressure, level of consciousness and urine output) that are measured and recorded based on an instruction that we have discussed in both papers. The PUP score would specifically provide the ICU specialist with an initial evaluation of the health condition of a candidate patient for the ICU. This initial assessment could be even more accurate for the MET call as a junior doctor also accompanies the PART nurses in assessing the illness severity of patients and consults with the senior ICU specialist about the patient’s health condition before any ICU admission. This information would be precious for the ICU to manage its capacity prior to making any admission or discharge decision. On the contrary, the severity of illness scores used in S.-H. Kim et al. (2014) are obtained from a lab, 24 hours preceding of a patient hospitalization, and an estimated probability of patient mortality.

In addition, we have presented an analytical approach to the problem of nurse staffing between PART and ICU. Although some studies have developed queueing models to address nurse staffing in a hospital (for example, see Armony et al. 2017; de Véricourt & Jennings 2008, 2011; Yankovic & Green 2011), we are not aware of any studies that proposed an analytical solution to this challenge in hospitals. In particular, we have sought to represent a realistic view of different patient flows including both false positive and false negative patients arriving to the ICU in the simulation model in Paper II. The proposed model provides an insight into how nurse allocation policies between PART and ICU could impact the ICU patient mortality rate.

We also believe that our study contributes to the medical literature. Despite the fact that the roles and efficacy of the outreach services on patient outcomes were discussed in-depth
in the medical literature, these studies lack an analytical approach toward analysing the
effect of nurse staffing in the ICU on the patient outcomes. In other words, most studies
relied only on prospective cohort or before-after studies while the evidence for failure of
the outreach services implementations in hospitals can prove that, a genuine and robust
interpretation of such services in hospitals will not be meaningful if the research only
counts on prospective studies. We have attempted to fill this gap in the medical literature
in Paper II.

Nonetheless, in the medical literature, the MERIT study carried out by J. Chen et
al. (2015) in Australia is best aligned with our research in Paper I. The MERIT study
specifically investigates the impacts of delayed MET calls on ICU patient outcomes in 23
hospitals in Australia. Their findings showed that for cardiac arrest patients, 15 minutes
delay after detection of the instability of the patient is independently associated with an
increased risk of ICU admission and death. In contrast to the MERIT study and most other
medical research that examined only the medical variables, we have concurrently included
both operational (e.g., the occupancy level of nurses) and patient-level variables in our
econometric models in Paper I. We are confident that our research took a new look at
the scenarios that reflect more of the variables involved in the current decision-making
processes in ICUs.

6.4 Contribution to Practice

According to Digital Universe Driving Data Growth in Healthcare (2014), healthcare
data is increasing at 48% per year, even faster than the rest of the Digital Universe
(https://www.emc.com). More surprisingly, it has been projected in the report that by
the year 2020, 50% of the health systems in the US will be on the second generation of
Electronic Health Record (EHR) technology. This is a unique opportunity for Business
and Data Analysts at hospitals to use available data to develop models that provide more
insights into the current challenges in hospitals. Despite the significant advantage of ex-
isting data analytic tools and software in analysing the day-to-day processes as well as predicting the future behaviour of the healthcare system, the decision-makers at hospitals are still challenged with the problem of resource allocation to different medical units. However, the simulation model in Arena, for example, can address most of these concerns once it is integrated with the data analytic tools. Modelling the uncertainties involved in the patient arrivals and service times to obtain the required number of nurses, doctors and beds is one of the advantages of the simulation study. Besides, the visualisation of the dynamic behaviour of the system even at the individual patient level could be very helpful for decision-makers (e.g., a house officer or a nurse coordinator at ICU) to evaluate the results of planned nursing policy on patient outcomes and ICU congestion level in advance. Finally, the “what-if” scenarios, instead of costly and sometimes even impossible experiments, could be very useful to managing the scarce nurses in the ICU. In Paper II, we have discussed some of these applications and quantitatively shown how the simulation model in the nurse staffing problem between PART and ICU could bring a considerable saving for the hospital. Therefore, we would recommend that hospitals use the benefits of simulation modelling in their decision-making process as it is not costly and can reduce the risk and consequence of unimplemented policies.
Chapter 7

Future Research Direction

We acknowledge that our study has some limitations that open a new avenue for the further research. First, we have validated our proposed model based on one-year data collected from a single hospital in New Zealand. The results could be more accurate if one validated our model using a larger dataset. Second, the PART that we studied is a nurse-led outreach service while other configurations of outreach services such as MET, RRT or CCOT with different compositions (e.g., a medically-led or a multidisciplinary team) have been implemented in hospitals throughout the world. Applying our approach, it would be interesting to find out what combinations of nurses and doctors would result in the best for both patients and hospitals. The trade-off between the costs of constructing a multidisciplinary team and team effectiveness in improving patient outcomes and ICU performance could be of interest to both researchers and practitioners. Third, in a recent study, Hu et al. (2015) used an Early Detection of Impending Physiologic Deterioration version 2 (ESPI2) that electronically measures and updates the illness severity of patients in the ward every six hours. As the applications of Electronic Health Record (EHR) among hospitals are increasing, one could use the same methodological approach and test how the results will be changed once ESPI2 is used instead of the PUP scoring system. In this regard, we think that our findings would have been more concise if we could have acquired more information on ward patient illness severity (e.g., lab data). Besides, a new EWS
system has been just introduced to New Zealand hospitals (see Figures 1.12 and 1.13). Future research can apply the proposed methodology to analyse the impacts of the new EWS system the patient outcomes and ICU admission policies. More interestingly, one could also explore the rate of false positive and false negative that the new EWS could generate. Fifth, we have not modelled the patient flow between HDU and ICU, assuming both units are a single unit. Future research could propose the optimal nursing among PART, HDU and ICU when the ICU and HDU are set up separately. Last but not least, more generalised assumptions than those that we have made in the queueing model concerning patients’ arrival and service time processes could be considered. For example, one could use a non-homogeneous Poisson arrival process.
References


doi: 10.1016/S0305-0548(03)00034-0

doi: 10.1023/B:ANOR.0000019103.31340.a6


Altiok, T., & Melamed, B. (2001). *Simulation Modeling and Analysis with ARENA*. S.l., Warren,


Angus, D. C., Linde-Zwirble, W. T., Lidicker, J., Clermont, G., Carcillo, J., & Pinsky, M. R.


Arabi, Y., Alshimemeri, A., & Taher, S. (2006). Weekend and weeknight admissions have the same outcome of weekday admissions to an intensive care unit with onsite intensivist coverage. *Critical Care Medicine, 34*(3), 605–611.


doi: 10.1002/net.3230200103


References


References


References


References

Electrophysiology of the European Society of Cardiology, 8(6), 449–455. doi: 10.1093/europace/eul042


Chua, W. L., Mackey, S., Ng, E. K. C., & Liaw, S. Y. (2013). Front line nurses’ experiences


team. Critical Care Medicine, 34(1), 211–218.


DoH. (2000). *Comprehensive critical care: a review of adult critical care services*


Duffin, C. (2014). Increase in nurse numbers linked to better patient survival rates in ICU. *Nursing Standard* (Royal College of Nursing (Great Britain): 1987), 28(33), 10. doi:

doi: 10.1177/0037549707083111


References

296–300.


Franklin, C., & Mathew, J. (1994). Developing strategies to prevent inhospital cardiac arrest: analyzing responses of physicians and nurses in the hours before the event. Critical Care Medicine, 22(2), 244–247.


Frost, S. A., Alexandrou, E., Bogdanovski, T., Salamonson, Y., Davidson, P. M., Parr, M. J., &


doi: 10.1287/msom.1110.0341

doi: 10.1214/aoms/1177728975


doi: 10.1016/S0305-0483(02)00050-6

doi: 10.1016/S0272-6963(00)00027-9

doi:
10.1016/S0377-2217(98)00135-0


Lane, D. C. (2000). *You just don’t understand me: modes of failure and success in the discourse between system dynamics and discrete event simulation* [Monograph]. Retrieved 2018-


Li, G., Thabane, L., Cook, D. J., Lopes, R. D., Marshall, J. C., Guyatt, G., … Levine, M. A. H.
References


doi: 10.1145/167293.167839
doi: 10.1177/003754979406300402


models the need for critical care resources. *Anesthesiology, 100*(5), 1271–1276.


GlobeEdit.


References

Routledge.


Zimmerman, J. E., Wagner, D. P., Draper, E. A., Wright, L., Alzola, C., & Knaus, W. A.
