ABSTRACT

NAMBIAR, SIDDHARTHA. Improving Hospital Operational Efficiency when Staff Workload Affects Patient Outcomes. (Under the direction of Maria Mayorga.)

The U.S. Institute of Medicine has estimated that approximately 44,000-98,000 patients die annually due to medical errors. A large proportion of these errors stem from inefficiencies in hospital operational decisions and poor utilization of resources. In this dissertation, we aim to improve hospital operational efficiency by developing a set of models that optimize different decisions within the health care process, such as patient flow and resource utilization. In developing these models, we propose to incorporate features unique to healthcare systems that make the models better abstractions of reality. These features include the use of nurse workload to affect patient outcomes, the use of floating nurses to handle abrupt variation in demand, and the pooling of resources within a hospital unit to care for the patients within the unit. We have divided this dissertation up into three bodies of research.

First, we develop a strategy to dynamically allocate floating resources within a set of inpatient hospital units. Consider patients being treated for a disease whose condition changes over time. Resources allocated to a patient influence disease progression and outcomes. Our goal is to allocate a limited number of floating resources to patients, depending on their health status, to maximize long-run health state utility. We formulate a Markov Decision Process (MDP) to allocate the floating resources in such a setting. Our experimental studies show that the optimal allocation policy is sensitive to the choice of model parameters and choice of the objective function. We also test the optimal "dynamic" policy against various heuristics that are easier to either implement or to compute. We find that an optimal "static" policy performs better than the other considered heuristics.

The second main body of research within this dissertation considers the flow of patients within an emergency department. We develop a multi-server queuing model that routes incoming patients to one of several different server pools in an Emergency Department (ED). We perform the routing such that we minimize the total time spent by the patient in the system and that we efficiently manage the workload experienced by the servers (nurses). Our results show that using information about the long-run workload experienced by the nurses when routing incoming patients leads to a reduction in patients' length of stay (LOS).

The final body of research extends upon the patient routing model with an ED by considering patients

of several different severity types. We model a multi-class many-server pooled queuing system where patients of different acuity levels receive care from one of several nurse pools, each comprising an ED unit. We assume that a patient's time in service is a function of the nurse workload (for which we use the nurse-patient ratio as a proxy) in their unit. Our objective is to reduce patient Length-of-Stay (LOS) and to control nurse workload by optimizing routing and nurse allocation decisions between units. First, we address the complexity of the queuing model control equations due to patients of multiple acuity levels present in the same ED unit with service rates that depend on patient acuity and unit workload. To do this, we approximate the queuing system via a deterministic fluid model to describe the control equations via their first order behaviors. Next, we formulate the optimization model using the control equations obtained from the fluid approximation. The results of the fluid optimization problem are input to a simulation developed with AnyLogic software to obtain performance measures of interest for the non-approximated system, such as patient LOS and unit workload. We use data from a hospital in North Carolina, USA, to estimate parameters such as arrival rates, unit capacity, and service rates. Our results (1) highlight the importance of accounting for nurse workload and service behavior in developing routing/staffing policies and (2) show that small changes to patient routing policies could lead to reduced patient LOS and better-balanced nurse workloads.

Improving Hospital Operational Efficiency when Staff Workload Affects Patient Outcomes

by
Siddhartha Nambiar

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Industrial Engineering

Raleigh, North Carolina

2020

APPROVED BY:

Julie Ivy	Yunan Liu
Osman Ozaltin	Maria Mayorga Chair of Advisory Committee

BIOGRAPHY

Siddhartha is a doctoral candidate at the Edward P. Fitts Department of Industrial and Systems Engineering at NC State University. Born and raised in the south of India, Siddhartha moved to the Buffalo NY to get a Master's degree after graduating with a Bachelor's degree in Mechanical Engineering from the Indian Institute of Technology in Guwahati. His research interests lie in applying simulation and stochastic processes towards the field of health systems engineering and healthcare policy. After completion of his doctorate, Siddhartha will work as a Research Fellow at the MedStar Health Research Institute in Washington, D.C.

ACKNOWLEDGEMENTS

I would like to begin by thanking my advisor, Dr. Mayorga. She has played an immense role in my development as a Ph.D. student. Despite joining NC State with a Master's degree under my belt, I had very little confidence that I could go out into the 'real-world' and perform well. I have often joked with people that the reason I joined a Ph.D. program is because I wasn't ready to go out into the 'real-world'. Today, I take great pride in my work and look forward to when I graduate so I can actively make a difference with my career. None of this transformation would have been possible without Dr. Mayorga's constant support and encouragement.

Next, I would like to acknowledge the National Science Foundation's (NSF) grant number 1522107 which provided me with monetary support to conduct this research.

The comments, suggestions, advice, and help provided by members of my dissertation committee have been valuable for me. I'm thankful to each and every one of them for their guidance.

Thanks are due for all of my friends and coworkers for providing me with much-needed camaraderie. I'm thankful for people like Nikhil, Vishesh, Lena, Rachel, Rachael, Breanna, Joe, and so many more for gaming with me, laughing with me, and commiserating with me during my time in Raleigh.

My extended family has played such a huge role in helping me maintain my mental sanity by allowing me to freeload off of them during family vacations and gatherings. I'm thankful for their love, their support, the guidance, and the laughter.

My fiancée Elizabeth and our two cats Carly and Stella are the best partners I could have asked for. My life, and consequently my work, has improved tremendously since their introduction into my life. I'm thankful to them for so much and more.

Finally, I'd like to thank my parents. I will never be able to express the extent to which they have supported me through life. Both my mom and my dad have made a number of sacrifices for me throughout their lives and for that I will be eternally grateful. Their constant faith in my ability has never ceased to astonish me and I will forever be grateful to them for pushing me to my limits, for their constant support, validation, encouragement, and love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

1.1 The Need for System-Level Reform in Healthcare

A part of the Hippocratic oath that every medical student takes is the phrase "I will abstain from all intentional wrong-doing and harm". The idea behind this phrase is that under no circumstance should a medical professional undertake an activity that leads to the harm of a patient. While the big-picture intent behind the phrase is clear, factors that could 'harm' a patient are complex and not always well understood. Let's consider the following hypothetical situation taking place in a hospital ward. A nurse is currently in charge of three different patients, all of whom appear to be ready for discharge. A physician soon comes around and determines that all three patients may be discharged and promptly authorizes it. At the outset, this seems like an ideal situation for everyone involved. However, soon after all the patients are discharged, 3 new patients are admitted into the ward under this nurse's care. All of a sudden, our nurse went from having no patients to having to deal with three new patients and their onboarding processes. This inadvertently leads to a sudden increase in workload experienced by the nurse. We can think of nurse workload broadly as the number of tasks a nurse needs to undertake on account of a variety of

factors related to the number of patients they are tending to. Studies have shown higher odds for adverse patient safety incidents when nurses experience higher workload levels [Sta04; Wei07; AK09] and as a result the new patients who have just been admitted in our hypothetical setting are at a higher risk for a patient safety incident.

This hypothetical situation described above leading to a higher risk of a patient safety incident often invokes two counter arguments. Firstly, that the nurse experiences a higher workload and that it leads to adverse patient outcomes may be debated by those who claim that a medical professional's primary role is to ensure patient safety and that stress/workload should not deter them from carrying out this role. Secondly, given that the nurse does experience a higher workload, the physician who authorized the patient discharges could have been more cognizant about the effects this would have on the nurse and in turn, on the patients. Unfortunately, both of these arguments stem from a perspective that attempts to blame the humans for their errors instead of designing a system that leads to fewer errors being committed.

The U.S. Institute of Medicine issued a report [Don00] in November 1999 about the consequences of medical errors that occur in hospitals. Titled "To Err is Human", the report led to an increased awareness of U.S. medical errors by concluding that between 44,000 to 98,000 people die every year as a result of preventable medical errors. Media response to the report suggested that the healthcare organizations were at fault [Don08]. However, what was lost during the short period in the frenzy that followed was the original intent of the creators of the report. Their intent was to break the silence surrounding medical errors by centering the conversation around trying to fix the healthcare system that is designed to induce mistakes instead of blaming the hard-working health professionals for the mistakes. The idea behind the report was to advocate for dramatic, system-wide reform.

A second report was later published by the Institute of Medicine in 2001 as a follow-up to "To Err is Human". This report, titled "Crossing the Quality Chasm" [Cor05] advocated for a fundamental redesign of the U.S. healthcare system and recommended improvements across six dimensions of health-care in the US: patient safety, care effectiveness, patient-centeredness, timeliness, care efficiency, and equity. Many of these center around the effective use of healthcare resources. For instance, the second dimension of effectiveness is centered around efficiently managing the overuse, underuse, and misuse of resources and services. The fourth dimension of timeliness calls for a systematic reduction in waiting time delays in the U.S. healthcare system for both patients and caregivers. The report highlighted the importance of

not blaming the healthcare professionals for the quality of care and insisted rather that “it is the processes in which the people work that lie at the root of our troubles”.

Many aspects of the healthcare delivery system are unique to the field of healthcare (e.g., personalizing care, nature of government interventions). However, similar problems may be found in other sectors like telecommunications, logistics, and production [Rya05] . Soon after the publication of "To Err is Human" and "Crossing the Quality Chasm", the Institute of Medicine partnered with the National Academy of Engineering (NAE) to explore whether perspectives from systems engineering could address the challenges being faced by the healthcare industry.

1.2 Health Systems Engineering and Key Drivers of a Health System’s Performance

Health Systems Engineering is an interdisciplinary field that incorporates aspects of engineering and management to design and maintain complex health systems. According to Schlager [Sch56], this idea of recognizing and modifying the properties of a system as a whole can be traced back to Bell Telephone Laboratories in the 1940s. Systems engineering has since been used in a wide variety of applications to improve the efficiency, quality, and safety offered by products and services. Fanjiang et al. [Fan05] advocated that tools used in manufacturing and production systems be translated into a language that applied to healthcare, especially since the argument could be made that the delivery of medical care is the most complex production challenge on the planet. The author continued on in his book to say that little production-process thinking was being used in healthcare and that the gap between medicine and engineering needed to be bridged. Over the past few decades, there has been an out-pour of healthcare professionals adopting systems engineering principles. Some common systems engineering concepts that have since been applied in the healthcare industry include project management models [End06], engineering economics [Ste06], statistical modeling [Par03], stochastic rocesses [Hau00], operations research models [Bre05], and discrete event simulation [Hun07].

A core tenet of systems engineering is identifying the most relevant system characteristics and abstracting them in the form of mathematical models. This abstraction of the original system is then studied in order to understand the workings of the original system and interventions are introduced into this abstraction to try to improve the performance of the original system. Hospitals may be thought of

as large and extremely complex systems. From a systems modeling standpoint, it is difficult to capture all of a hospital's workings into a single detailed model and as a result researchers have tackled various subsets of hospital operations and tried to improve the performance of the subset they study. At a holistic level, the hope has been that improving the performance of a few parts at a time would lead to an overall improvement of a hospital's efficiency. Over this dissertation we approached the problem of improving the efficiency of a hospital by considering its subsystems. We explored resource allocation within an inpatient setting in chapter 2, and in chapter 3 we modeled patient flow within an emergency department with a focus on controlling the workload experienced by nurses. Chapter 4 extends the work done in Chapter 3 by considering an emergency department setting with patients of multiple severity types. Each chapter thus looks at improving the operational efficiency of a different segment within a hospital's operations. To develop and describe these models, we used tools like markov chains, queuing processes, birth-death processes, fluid approximations, and discrete-event simulation. These tools can be classified under the broad umbrella of stochastic processes. Stochastic models and processes are excellent tools to represent real-world processes, phenomena and activities and are used extensively in health systems engineering research [KK07; Hal12; Gre05; Gre13].

Finally, this dissertation considers a set of unique salient features that have not typically been modeled in health systems engineering research. The following section outlines these features and provides an overview of historical background behind each one. Furthermore, we discuss how we operationalize them within the settings considered in Chapters 2, 3, & 4.

1.3 Unique Features of Healthcare Operations Modeled in This Dissertation

Over this dissertation, we will study and model different hospital subsystems (inpatient units and emergency departments) with the goal of improving metrics of efficiency within each subsystem. Additionally, we also consider and model unique features pertinent to healthcare operations (floating nurse pools, pooling of hospital resources, and nursing workload) that introduce realism into our models. Each of these features is discussed below.

1.3.1 Floating Nurse Pools

In nursing, “floating" refers to moving between units. Floating nurse pools thus refer to a dedicated group of nurses who are are able to move between hospital units as required [PL11]. Some hospitals have dedicated float nurse pools while others require permanently assigned nurses to act in a floating capacity because of staffing needs. The use of floating staff began as a means of improving staffing efficiency by using an available nurse from a different unit undergoing reduced levels of workload to assist nurses in a unit with higher workload levels. As this flexible staffing technique gained popularity, nurses began to show signs of dissatisfaction caused by stress and anxiety due to lack of sufficient cross-training provided to them before being asked to float [Hen5]. An essay published in MedSurg Nursing by Bates [Bat13] found that nurses who float to unfamiliar units could end up feeling unproductive as their time was usually spent searching for supplies or seeking assistance from the other nurses on the unit. This combination of dissatisfaction, stress, and anxiety experienced by floating nurses was also shown to impact patient safety [Hen5].

In response to this dissatisfaction among floor nurses required to float on an as-needed basis, hospitals across the country started creating dedicated teams of full-time float nurses [Pya15]. For instance, MedStar Washington Hospital Center has a float pool composed of 91 nurses[Pya15]. Further, the float pool is comprised of nurses who willingly want to pick up shifts on their off days. Some hospitals like Aultman Hospital in Canton, Ohio have even eliminated mandatory floating for floor nurses and ensure that floating is performed by the dedicated pool of float nurses [Goo11].

The use of float nurses in health systems engineering models is not new. Past researchers have incorporated float nurses in their models as a means of coping with abruptly varying demand levels [Wan12; Mae13; Bor99; Her74; Kao81]. While some researchers like Kao & Tung [Kao81] studied dedicated float nurse pools, others like Hershey et al. [Her74] approached float nurses as floor nurses who may be re-assigned depending on demand requirements. Researchers like Maenhout & Vanhoucke [Mae13] have also incorporated a penalty representing ’float nurse discomfort’ to model the disadvantages that may arise from floating a floor nurse. A more comprehensive review of literature regarding the use of float nurses in engineering models is provided in Chapter 2 of this dissertation, which deals with the dynamic allocation of a dedicated pool of floating nurses.

1.3.2 Pooled Resources Within Hospital Wards/Units

In modern-day healthcare practices, patients are rarely looked after by just one health professional. Gone are the days when patients were cared for by a single physician or a private nurse residing in the community [Mit12]. Modern healthcare is complex and has evolved such that health professionals work as members of teams who share common aims with respect to patient outcomes [Mit12].

Traditional models of queuing theory [Gro08] deal with service systems in which a customer is looked at by exactly one server. Models incorporating multiple servers tend to assume that each one serves a different customer. When there are more customers than servers, there typically exists a waiting room within which customers wait until a server is available to see them. A variation of this type of model is the “Processor Sharing" service discipline [Dud18]. Here, customers are all served simultaneously with each receiving an equal fraction of the available service capacity. Processor sharing models have been applied most extensively in the field of multiprogramming computer systems.

In this dissertation, we contend that the service being provided by a pool of nurses in a hospital unit to the patients within the unit is similar to that of a “processor sharing" service discipline. In both Chapters 2 and 3, we assume that adding or reducing more patients to a unit changes the service rates of all the patients within the unit. As we discuss in the next subsection, we incorporate this by modeling a patient's time in system using an exponential distribution and having its rate be a function of the nurse-patient ratio in the unit.

1.3.3 Nurse Workload and Patient Outcomes

Nursing workload is an important factor in the evaluation of operational procedures (e.g. staffing decisions [Upe07]) in healthcare systems. While the idea of nursing workload is broadly related to the number of tasks a nurse needs to complete, a universally accepted definition for nursing workload does not exist [Car08]. Different researchers and healthcare organizations use a variety of ways to define or calculate the workload experienced by the nurses. Some use the severity level of a nurse's patients to calculate the workload [Lan04] while others have the nurses fill out questionnaires to evaluate their workload subjectively [Oet16]. Later in this section and in each chapter of this dissertation, we have discussed our method of modeling nursing workload using the ratio of nurses to patients within a hospital ward.

One of the earliest mentions in the literature about nurse workload affecting the operational perfor-

mance within a hospital was undertaken by the United States Air Force. The purpose of this study by Loschiavo [Los89] was to compare the ways in which nurse staffing requirements are determined in Air Force and civilian hospitals. The study noted that Air Force hospitals used a system called workload management system for nurses (WMSN). This system was based on estimating the amount of direct and indirect nursing care that would be required for patients in order to establish staffing levels.

Several studies have shown that higher nursing workload has adverse effects on patient safety [Car08; AK09; Coh99]. Aside from being correlated with sub-optimal patient care, high levels of nurse workload have also been shown to lead to reduced patient satisfaction [And98]. Much of the literature exploring how nurse workload affects patient outcomes has involved studying how nurse staffing affects patient outcomes for which there is strong evidence of correlation [Gri16]. Quantifying nurse workload is by no means an easy task, as workload is a complex construct. A commonly used measure to represent nurse workload is nurse-patient ratio [Car08]. One reason for the extensive use of this measure is because it is easy to use and available in databases and electronic hospital records. Some researchers believe that mathematical models that use nurse-patient ratio as a workload measure do not offer a significant contribution to capturing the impact of nursing workload in their models [Car08]. However, there is a lack of a consensus on how to objectively quantify workload as a composite measure. This lack of consensus coupled with an increasing need for quantifying objective workload measures using data available in electronic health records leads to surrogate metrics such as 'nurses per occupied bed' [Twi09], 'nurse to patient ratio' [Hur08], 'ward design and size' [Hee95] and 'high variety of case mix' [Sou00] needing to be used in developing models to improve operational efficiency.

This dissertation uses nurse-patient ratio as a surrogate for nursing workload and as one of the driving factors that affects how a patient's health condition changes. All of the three upcoming chapters have assumed that the time spent by a patient in a unit is distributed exponentially with the mean being a function of the nurse-patient ratio in the unit. We will see in the results section of both chapters how workload plays a role in operational decisions.

1.4 Document Outline

This section provides an outline of what to expect over the upcoming chapters. The over-arching goal of this dissertation is to improve the operational efficiency within some sectors of a hospital while

taking nursing workload into account. In Chapter 2 we look at an inpatient hospital system with a pool of floating nurses that need dynamic re-assignment between the hospital units depending on the distribution of patients within the units. The time spent by the patients in each of the units is modeled as a function of the ratio of nurses to patients within the unit. The form of this function depends on whether the patient's condition is improving or deteriorating. Each time a patient enters or leaves a unit, a hospital administrator has the option of redistributing the floating nurses. The objective of this study is to determine an allocation strategy that maximizes a measure of long-term reward, defined according to the time spent by a patient in the system and patient discharges (both favorable and unfavorable). The results show that the allocation strategy depends not only on the form of the function relating nurse-patient ratio and average patient time in unit, but also on the choice of how to model the objective function. Specifically, the choice of whether we wish to maximize the favorable patient outcomes or minimize the unfavorable patient outcomes leads to significantly different allocation strategies, under certain conditions.

While Chapter 2 discusses staff allocation within inpatient hospital units, Chapter 3 attempts to improve the efficiency of patient flow through an emergency department. Like in Chapter 2, we assume that the mean value of the distribution governing a patient's length of stay within an ED ward depends on the workload experienced by nurses within the ward. We consider nurse-patient ratio as a surrogate for a workload metric. The goal of this work is to develop a strategy to route incoming patients to an appropriate ward so as to minimize the average time spent by a patient in the system. Additionally, we also include constraints within the model that forces the average workload experienced by the nurses within a unit to stay under a predefined threshold. Results from this chapter show that always assigning an incoming patient to the ward expected to have the least increase in workload on patient assignment does not perform the best. We show the existence of a routing policy that that takes into account other system-level factors, such as patient arrival rate, to route patients that performs the best.

Our model in Chapter 3 focuses on patients of a single severity type. While we provide a short discussion at the end of the Chapter about how the framework could be extended to multiple patient types, a rigorous treatment is omitted. However, Chapter 4 extends upon the modeling premise established in Chapter 3 by looking at the flow of patients of multiple severity types within an ED. Like in earlier chapters, Chapter 4 considers nurse workload to be an integral factor that determines patient service time by modeling the average patient time in service as a function of the nurse-patient ratio within

a ward. An important addition to this Chapter is the use of real hospital data from the Emergency Department of Southeastern Regional Medical Center located in Lumberton, North Carolina. We show the existence of a relationship between patient time in service and nurse-patient ratio before characterizing the relationship and using it to develop a queuing model. We then discuss the difficulty involved in quantifying the equations for this complex queue involving state-dependent processor-sharing style service rate and develop an approximation using the concept of fluid queues. In addition to developing fluid approximations, we create a simulation of the ED to test out the solutions from the fluid approximation. Our objective in this chapter is to re-design patient routing and nurse staffing policies that reduce patient LOS and better manage the workload of nurses within the wards. Our results show that this objective is achievable with small modifications to existing policies. Furthermore, our work establishes a framework that may be implemented within a real-time clinical decision support framework using a combination of fluid approximation to queues and simulation modeling.

Chapters 2, 3, & 4 are each formatted as stand-alone essays and discuss three separate projects. Each Chapter begins with an introduction about the project followed by a review of relevant literature before outlining the mathematical model and discussing related results. Finally, in Chapter 5 we tie the three projects together and discuss the implications of the research done throughout the dissertation. Chapter 5 also outlines a series of possible new projects and extensions to the existing projects that could advance the work done in this dissertation.

CHAPTER

2

RESOURCE ALLOCATION STRATEGIES UNDER DYNAMICALLY CHANGING HEALTH CONDITIONS

2.1 Introduction

According to the Centers for Medicare and Medicaid Services [Cms], the US national health expenditure has increased by over $1 Billion between 2007 and 2017. Rising healthcare costs affect the entire healthcare system, including patients, providers and payers. One way to mitigate these costs is by increasing the efficiency of healthcare delivery. The efficient allocation of resources plays an important role in ensuring that patients receive the best quality of care [Bar16]. Like with most service systems however, such an allocation involves trade-offs. For instance, an over-provision of hospital beds leads to sub-optimal bed occupancy [Goe14]. Conversely, an under-provision of hospital beds results in patients requiring admission being denied appropriate medical care. While the relationship between resource

allocation strategies and patient outcomes has been widely studied/accepted, traditionally resource allocation in healthcare follows fixed allocation strategies. However, it has been seen in healthcare and in general service systems that flexibility in the management and allocation of resources decreases variability and increases efficiency [Ver09; Whi06a]. Thus, using flexible resources is one way that hospitals can increase efficiency of care.

In this study, we consider allocation strategies for floating resources, such as floating nurse pools, in healthcare settings. Floating nurse pools consist of nurses who are trained to work in multiple units throughout the hospital, thus allowing the hospital to adapt to the ever fluctuating needs and volume of patients [Leb15]. The use of these floating nurse pools and resource teams (like Rapid Response Teams (RRTs) [Sto15]) is also often a strategy used in hospitals to cope with variable or inadequate staffing. Floating resources are flexible and are assigned when a unit's census is high, or to fill staffing shortages across the hospital due to absences, vacancies, or high acuity levels. Such resources benefit healthcare organizations [Gos98; Cav02] and have been found to be effective in reducing staffing costs in inpatient settings [DE06; PL11].

Unlike hospital beds, human resources like nurses and other hospital personnel (say, members of the team that make up an RRT), are not allocated one-to-one to patients. This is because a significant portion of the care process within a unit often involves collaboration between the personnel present within that unit. As a result, patients are thought of as being assigned to a unit or ward instead of being assigned to a single nurse. In this vein, the patient's length of stay depends not only on the complexity of their own condition but also on several other factors that that are related to their unit such as shift exchanges, lack of a standardized process of admission, lack of sufficient staff members, communication failures, number of other patients, and overcrowding [Bas15].

In this chapter, we consider a healthcare system where we can assign floating resources among units comprised of patients at different severity levels. We present this as an adaptive resource allocation model using a Markov Decision Process (MDP). On entry into the system, patients are assigned to a unit based on the severity of their condition. As their severity changes, patients are transferred between units. Armony et al. [Arm13] describes an operational example of such a system in which they use step-down-units to provide an intermediate level of care between intensive care units (ICUs) and the general units. In this chapter, we model a health system with multiple units housing patients of different severity levels. Each unit is staffed by a set of permanent (or baseline) nurses in addition to a set of

floating nurses who transition between the units. The systems manager decides how many floating resources to assign to the cluster of patients in each unit. We characterize the state of the system by the number of patients in each unit. We assume that a patient's severity improves or deteriorates depending on the number of resources allocated to the unit that the patient is in. We implement this by modeling the patient's transition between different severity levels (between different units) as a function of the ratio of patients to resources. To do this we assume that the time until a patient transitions is exponentially distributed with a state-dependent rate. We test the model with different functional forms for this rate (linear, convex or concave) to see how it affects the resource allocation strategies. In other words, the average amount of time (mean value for the exponential distribution) it takes for a patient to transition to a new state is modeled as a linear, convex, or concave function of the total number of patients and staff in the ward that the patient is in. The choice of functional form represents how we believe marginal increases to the workload affect patient outcomes (represented in this article via patient length of stay (LOS)) within a unit. For instance, if the time a patient spends in a ward is a convex function of the nurse-patient ratio of the ward this implies that for a patient whose health condition is improving, the marginal decrease in the time they have to spend in the unit decreases with an increase in the nurse-to-patient ratio. In other words, a patient's health condition improves slower at higher nurse-to-patient ratios indicating that the nurse productivity decreases at higher workload levels. Concave and linear functions show that the marginal decrease in time spent in the unit decreases or remains the same, respectively as nurse-to-patient ratios increase.

We provide results for optimal allocation policies which we then compare against heuristics to explore the trade-off between ease of implementation and optimal performance.

2.2 Literature Review

We begin this literature review by summarizing resource allocation techniques in both healthcare and other general service systems. For health-related literature, we restrict our review to human resources in secondary healthcare settings, such as acute care (found in emergency departments). Examples of human health resources include physicians, nursing professionals, and hospital technicians. We do not review literature related to scheduling and instead focus on staffing capacity.

For literature related to resource allocation in general service systems, we focus on queuing systems

that involve one or multiple customer arrival classes with state-dependent service rates. We review those papers that explicitly include allocation policies for resources, in this case, determined to be the servers of the queuing system, as one of the primary research goals. We then outline gaps in the current literature and discuss our contribution.

2.2.1 Resource Allocation in General Production/Service Systems

The question of determining optimal staffing/resource allocation has been studied extensively over the years. Moses [Mos72] considered the problem of allocating and dispatching servers in a setting where the system could experience random failures and would require repairs at the point of failures. The author provided examples for systems like telephone or electric lines, train tracks, highways and gas transmission lines.

A commonly studied staffing policy considered by recent literature is called the square-root staffing policy [Bor04]. When there are linear waiting and staffing costs, the authors show that to achieve economic optimality, the system needs to employ a number of servers equal to the sum of the offered load (ratio of effective arrival rate to total service rate) and the square root of the offered load. Other researchers have shown that this staffing policy is robust and its optimality continues to hold even in cases of added complexity, such as customer abandonment [Gar02] and uncertain arrival rates [Koç15].

Finally, we note the work of Gurvich & Whitt [Gur09] and Mandelbaum & Stolyar [Man04] who considered scheduling flexible servers under different settings. The former [Gur09] considered the case of many-server service systems with multiple customer classes and server pools while the latter [Man04] considered a queueing system with multitype customers and flexible (multiskilled) servers that work in parallel. We note here that our work differs from the above in that we consider the allocation of resources by addition or removal of servers from the server pool and not by routing customers/patients to specific servers/server pools.

2.2.2 Resource Allocation in Hospital Settings

We find one of the earliest mentions of health-related resource allocation techniques in the work of Zemach [Zem70]. The mathematical model presented in their paper considered the utilization of personal health and medical services by the population of a community or region. Using past data, the author predicted resource use and allocation over time as a function of population dynamics.

Green et al. [Gre06] and Green & Kolesar [Gre04] provide an overview of more recent literature relating to hospital resource allocation techniques. A common resource allocation problem that has been considered by the operations research community is related to staffing physicians [Lip98; Liu18; Sir17] and nurses [Hel80; Gre13; Vér11]. These issues have been addressed using a number of tools including mixed-integer programming [Hel14; Wan14], queueing theory [Gop16; Liu18], and simulation modeling [Sir17]. Recent literature also comprises of models that incorporate more realistic features like time-varying demand [Che09], multi-period planning [Car16], staff workload [Sim11], and multidisciplinary teams [Ago17].

2.2.3 Distinguishing Features of Our Model

The various complexities of healthcare make it difficult for any one model to be an accurate abstraction of reality. In our work we focus on three realistic features of a hospital system to model analytically, namely demand-driven staffing, pooling/sharing of resources within a ward, and staffing-dependent lengths of stay for patients. Below we describe the literature and relevant gaps surrounding these features.

2.2.3.1 Demand Driven Staffing.

Green et al. [Gre06] and Green & Kolesar [Gre04] noted that much of the previous literature has assumed fixed staffing levels needed for each shift and provided one of the first queuing models to guide staffing decisions. Two common staffing models used by hospitals today are staffing by nurse-to-patient ratio and budget-based staffing. An overview of these staffing models is provided by Mensik [Men14]. Both aforementioned staffing models use information about average occupancy levels to determine how to allocate nursing resources. The use of average occupancy levels to allocate resources and manage capacity is unreliable from a managerial standpoint. Green [Gre05] provided various reasons for this, some of which are that the time when hospitals count patients for billing purposes is typically at midnight, that utilization of hospital facilities during a week is non-uniform, and that reported occupancy levels are usually based on yearly averages and do not account for temporal trends.

We examine in our work this need for demand-driven staffing by considering the use of floating nurse pools. In the work by Warner [War76], we find one of the earliest mentions in the literature detailing the use of floating nurse pools in analytic models. Warner discusses 'fine-tuning' the nurse scheduling decisions by allocating a pool of available floating nursing personnel to account for unforeseeable

variability in demand. In an annotated bibliography, Ernst et al. [Ern04] discussed computational methods for rostering and personnel scheduling; this included a set of papers [Abe73; Her74; MR73; Tri76] that incorporated the use of floating nurse pools to account for sudden variation in demand.

2.2.3.2 Pooling/Sharing of Resources Within Wards.

Few analytical models for resource allocation in healthcare consider the fact that resources within wards are partially shared, central resources [Sch13]. In general service systems, the use of pooled resources is related to the concept of 'processor sharing' [Kle67]. Processor sharing is a service policy where customers are all served simultaneously in a queuing system. Under processor sharing, each customer receives an equal fraction of the service capacity available.

Sharing resources within a ward is an idea that is relatively new in healthcare analytics literature. Agor et al. [Ago17] developed a simulation model in which incoming patients are assigned to teams of providers of different skill levels. Mandelbaum et al. [Man12a] showed that based on empirical hospital data the Inverted-V queuing model best models patients spending time in units within a hospital. The Inverted-V model assumes that upon entering a queuing system, an agent (patient) is assigned to a 'pool' of servers instead of being assigned to a single server. Several authors continued to build on this by proposing a variety of patient/customer routing algorithms in an Inverted-V queueing context [Alm13; Arm10; War13].

In addition to implementing float nurse pools to satisfy demand-driven staffing and considering adaptable lengths of stay (discussed in the next subsection), our model builds on the literature related to the pooling/sharing of resources by assuming that upon entry, a patient is assigned to a pool of nurses within a ward.

2.2.3.3 Staffing dependent Lengths of Stay.

A third key component that is missing from most analytic models related to hospital staffing are factors that account for patient length of stay. Schmidt et al. [Sch13] addressed the feasibility of a computer-supported bed management system wherein a patient's expected length of stay is dynamic and is affected by factors related to the overall state of the ward. While there is a large volume of literature that discusses various factors that affect a patient's length of stay [Lav76; Bur91; Gru06; Cor03; Yoo03; Fei77], we focus on the correlation between nurse staffing and patient health outcomes. A number of studies have

suggested that a higher nurse-to-patient ratio leads to a reduction in length of stay [Thu07; But11; Kan07; Sch15; Twi11]. Nurse-patient ratio has often been used as a surrogate to describe the workload being experienced by nurses [Hur08; Car09]. The primary challenge with incorporating a dynamic LOS metric into models has been the lack of literature that analytically describes how nurse staffing affects patient health outcomes. Pitkäaho et al. [Pit15] used Bayesian dependence modeling to analyze the relationship between many variables (including patient acuity and the proportion of registered nurses (RNs)) and LOS. One of their key findings was that the relationship between the proportion of RNs among the entire nursing workforce and the patient's length of stay was non-linear and that above a limit, increasing the proportion of RNs was unable to predict if patients would have short hospital stays.

Most of the models discussed in this literature review include strong assumptions that restrict their use and often do not consider or cannot be adapted to incorporate a variety of real-world issues (like nursing shortages [Fri05] and provider workload [Car08]). In our work, we direct our attention to demand-driven staffing, pooling/sharing of resources within a ward, and staffing-dependent length of stay for patients and propose a model based on these features. Specifically, we develop a resource allocation model for floating nurse pools that recognizes that patients are assigned to the entire unit and not a specific nurse or staff member within a unit. Furthermore, a patient's length of stay depends not only their own health conditions but also on factors associated with the rest of the unit. In **??**, we conceptualize our model and develop an analytical formulation.

2.3 Model Description

We begin by providing a general overview of the system in Fig. 2.1. Patients arrive into the system with exponentially distributed inter-arrival times with rate Λ. These patients belong to K different severity types and each patient has a probability $p_k \; \forall \; k \in \{1,2...K\}$ of belonging to severity type k. Patients of each severity type thus arrive at a rate of $\lambda_k = \Lambda p_k \; \forall \; k \in \{1,2...K\}$. We define $n_k \forall k \in \{1,2,..K\}$ as the number of patients of severity type k. The state of the system, therefore, is $\bar{n} = (n_1, n_2, ..n_K)$. Upon entering the system, patients are assigned to a unit based on their severity type, such that a unit k, only holds patients of severity type $k \in \{1,2,...,K\}$. We allow a patient's condition to change over time, thus we denote $s_k \in S$ as the patient health state, where s_k refers to the health state of patients of severity type

k. Each unit k is assumed to have a baseline number of resources c_k allocated to it. Once assigned to a unit, a patient receives service in accordance to their severity type (further details regarding service rates are provided in Section 2.3.1). Upon completion of service, patients can leave the system into one of two states: routine discharge (recovery) or discharge to hospice (death). We use these two states as surrogates for desirable and undesirable patient outcomes respectively and may be substituted by other patient disposition types (such as discharge to a skilled nursing facility or rehabilitation center). To capture this mathematically, we create two absorbing patient health states $\hat{S} = \{h, d\}$ to represent recovery and death respectively. Note that for the rest of this article, we will use the concept of 'unit' and 'group of patients of a particular severity type' interchangeably. Our goal is to determine the number of floating resources c_k^f that need to be allocated to the respective unit (and consequently to patients of the respective severity type).

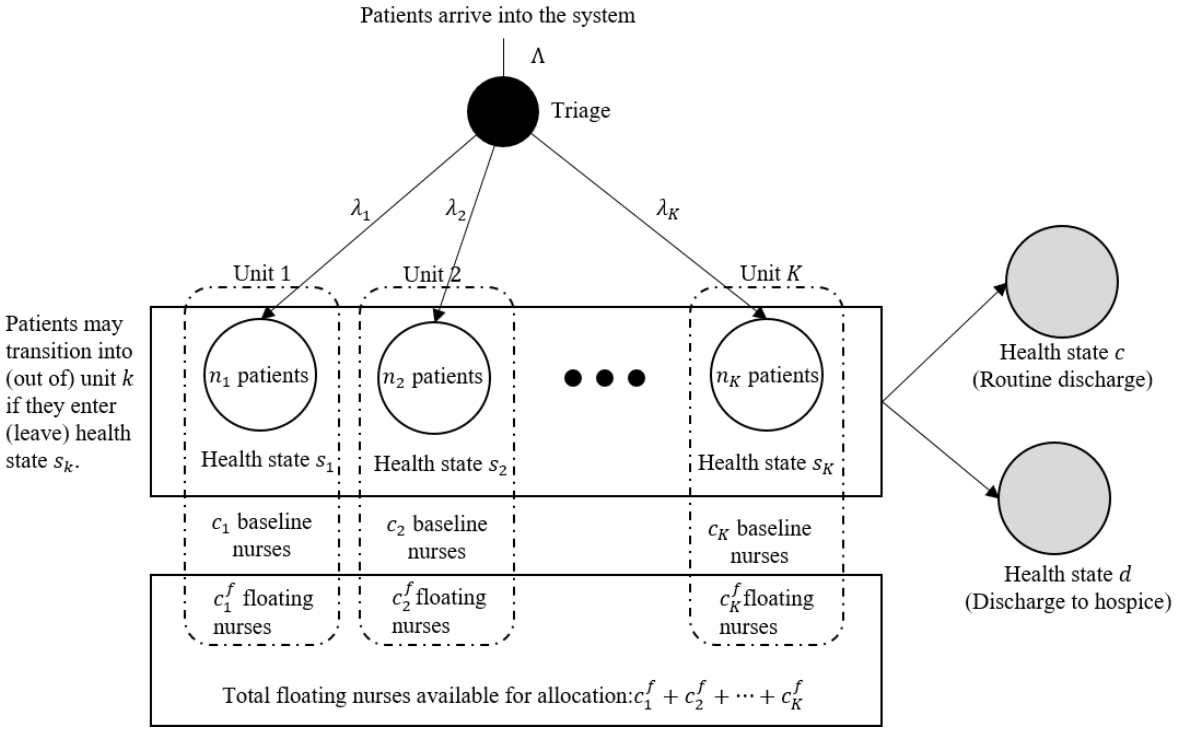


Figure 2.1 An overview of patient arrival and allocation process in the system: Patients on entry are assigned to units that correspond to their health state. Patients may transition in and out of a unit if they enter or leave a health state (patient health state transitions are not shown here). The objective of our model is to allocate $(c_1^f, c_2^f, .., c_K^f)$ floating nurses to units $(1, 2, .., K)$ having $(n_1, n_2, .., n_K)$ patients in health states $(s_1, s_2, .., s_K)$.

For any patient, let the set of all allowable transitions between the units contained in S be denoted by

E and the set of all allowable transitions from units contained in S to the absorbing states in $\hat{S}$ be denoted by $\hat{E}$. Fig. 2.2 shows an example of the state transitions from a patient's point of view when $K = 3$. A patient can be in one of three distinct health state ($k \in \{1,2,3\}$) as a result of which there exist exactly three distinct units $S = \{s_1, s_2, s_3\}$ in the hospital; the three units can be thought of as being the general unit, step-down, and ICU respectively. Here, it is assumed that a patient cannot go directly from unit s_3 to s_1, (i.e., from the ICU to the general unit). Similarly, it is assumed that a patient may only recover from unit s_1 (i.e., general unit) and that a patient may only die or be discharged to hospice from units s_2 and s_3 (step-down unit and ICU respectively). Thus in Figure 2, $E = \{(s_1,s_2),(s_1,s_3),(s_2,s_3),(s_2,s_1),(s_3,s_2)\}$ and $\hat{E} = \{(s_1,h),(s_2,d),(s_3,d)\}$. These transition times are exponentially distributed with an average value of $1/r_{ij}, \forall\{i,j\} \in E, \hat{E}$. The rate ($r_{ij}$) is a function of the number of patients in that particular unit and the number of floating resources assigned to it. Section 2.3.1 discusses this in further detail.

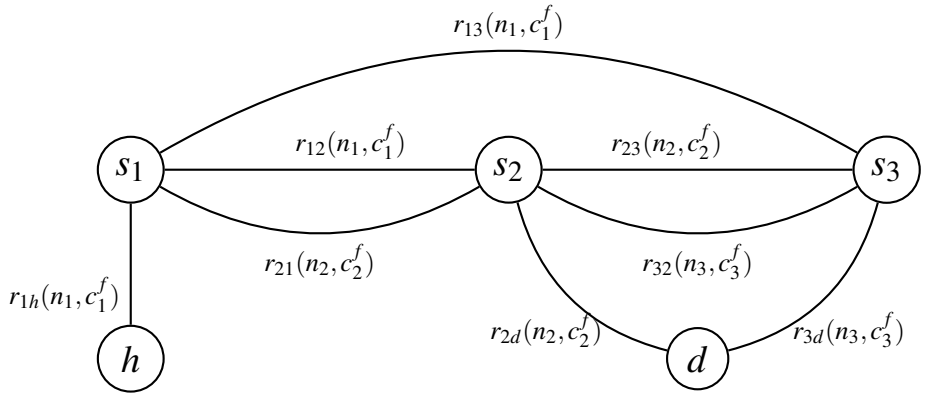


Figure 2.2 State transition diagram from an individual patient's point of view

2.3.1 Transition Rates and Utility Function

We define a stage of the process as the period during which the markov chain is in a particular state. The rate at which the system transitions to the next stage depends on the actions taken in the current stage. In our system, the transition rates between system states depend on individual transition rates; and the individual transition rates depend both on the number of patients of each severity type within the unit the patient is in and the number of floating resources assigned to them. Specifically, we assume that a patient's time in a health state is exponentially distributed with a mean that varies and depends on

the nurse-patient ratio (ε_i) of the ward i that the patient is in. In defining the average patient transition time function this way, we assume that the the resources allocated to a particular unit are shared by all patients present in the corresponding unit. Further, the average transition time function behaves differently depending on whether the patient's condition is improving or deteriorating. We define the functional form for the rate of transition function as

$$\mathrm{r}_{ij}(n_i, c_i^f) = \begin{cases} 1/(b_{ij}^{min} + (b_{ij}^{max} - b_{ij}^{min})q\left(\frac{\varepsilon_i - \varepsilon^{min}}{\varepsilon^{max} - \varepsilon^{min}}\right)), & \text{if } j > i \text{ or } j = d \text{ (i.e., deterioration)} \\ 1/(b_{ij}^{max} - (b_{ij}^{max} - b_{ij}^{min})q\left(\frac{\varepsilon_i - {}^{min}}{\varepsilon^{max} - R^{min}}\right)), & \text{if } j < i, \text{ or } j = c \text{ (i.e., improvement)} \end{cases}. \tag{2.1}$$

Here, b_{ij}^{max} is a baseline value representing the amount of time it takes a patient to leave unit i for unit j when the nurse-patient ratio ($\varepsilon_i = \frac{n_i}{c_i + c_i^f}$) is the maximum possible (ε^{max}) and b_{ij}^{min} is a baseline value representing the amount of time it takes a person to leave unit i for unit j when the patient-nurse ratio (ε_i) is the minimum possible (ε^{min}). Furthermore, $q(\cdot)$ is a function determining the transition rates. While running experiments, we will test different functional forms (convex, concave, and linear) for the transition rate function and will demonstrate how the chosen functional form affects the allocation of resources. The differences in functional form are representative of the differences in how nurses within a ward perform as their workload (represented via nurse-patient ratio) increases. A convex form, for instance, means that the marginal productivity of nurses decreases at increasing workload levels. A concave form, however, means that the marginal productivity of nurses increases at higher workload levels. In Section 2.4 we conduct numerical experiments to test the impact of these assumptions on the optimal allocation policy.

While the transition rate functions are defined from an individual's point of view, the system utility function is defined from a system perspective by computing the total health state utility value. These are commonly used as a component of quality-adjusted life year (QALY) calculations in population health and economic studies [Whi10]. The health state utility values attempt to reduce multi-dimensional health outcomes to a single representation or measure of health by being weighted against the average amount of time spent in a particular health state [Mue16]. They are commonly used in analytic models to study the benefits of treatments and interventions by representing the value of being in specific states of health [Ara11].

Let $U(\bar{n})$ be the total health state utility accumulated while the system is in state $\bar{n}$. Defining $u_1, u_2, ..., u_K$ as the individual utility values associated with having a patient in health state $s_1, s_2, ..., s_K$, we aggregate the health state utility values over all the patients across all units as

$$U(\bar{n}) = \sum_{i=1}^{K} u_i n_i.$$

Additionally, we define $U'_{ij}\ \forall (i,j) \in \hat{E}$ as the fixed cost or reward associated with the discharge to hospice or routine discharge of an individual, respectively.

2.3.2 Recursive Optimality Equation

In order to prepare the problem as an MDP, we first need to prepare the state transition diagram from a system's point of view. As described earlier, $\bar{n}$ is the state of the system. Fig. 2.3 shows possible state transitions for the system corresponding to the individual transitions described in Fig 2. Here, e_i denotes the K-dimensional vector whose i^{th} elements is 1 and the rest are 0. Accordingly, $\bar{n} - e_i + e_j$ implies that a patient left unit i to go to unit j, $\bar{n} + e_i$ implies that a patient entered unit i, and $\bar{n} - e_i$ implies that a patient from unit i left the system.

In determining the optimal allocation strategy, we formulate the problem as a markov decision process with the goal of maximizing the long-run expected utility. To solve the MDP, we use an equivalent Discrete Time Markov Decision Process (DTMDP) for the aforementioned CTMDP with a finite state space. We argue in favor of a finite state space by placing an upper bound on the number of patients that can be present in a particular health state. This is a reasonable assumption, since units have a limited number of beds and can typically not admit patients beyond a certain capacity. For the rest of this study, we will assume that each ward can have a maximum of M patients. This DTMDP process can be found by uniformization and discretization of the initial process [Lip75]. We formulate the optimality equation for the infinite-horizon average reward following uniformization as

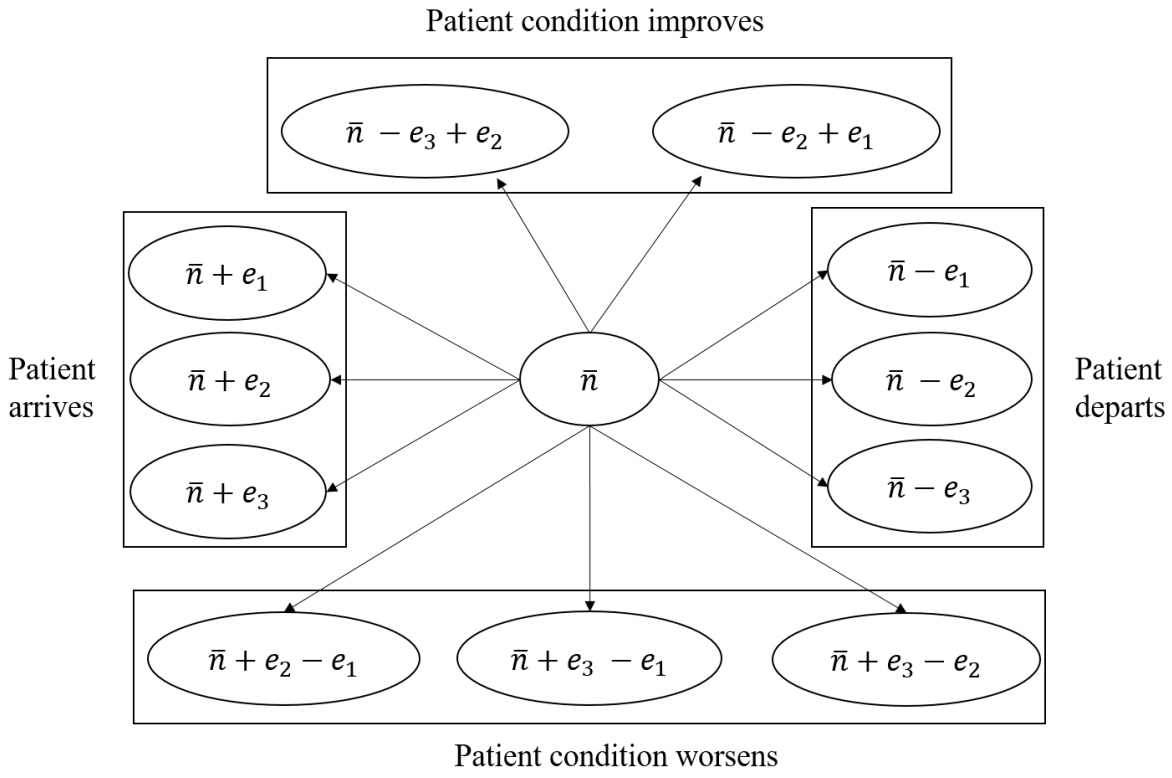


Figure 2.3 Possible state transitions from the system's point of view corresponding to individual patient transitions as described in Fig. 2.2. Note that e_i is a vector of zeros with the value 1 at the i^{th} position.

$$g + W(\bar{n}) = \frac{1}{\nu} \max_{\bar{c^f}} \begin{cases} U(\bar{n}) + \\ \sum_{i \in S} \lambda_i W(\bar{n} + e_i) + \\ \sum_{(i,j) \in E} n_i r_{ij}(n_1, c_i^f) W(\bar{n} + e_j - e_i) + \\ \sum_{(i,j) \in \hat{E}} n_i r_{ij}(n_1, c_i^f)[(\beta + \nu) U'_{ij} + W(\bar{n} - e_i)] + \\ \left(\nu - \sum_{i \in S} \lambda_i - \sum_{(i,j) \in E} n_i r_{ij}(n_1, c_i^f) - \sum_{(i,j) \in \hat{E}} n_i r_{ij}(n_1, c_i^f)\right) W(\bar{n}) \end{cases} \tag{2.2}$$

where g is the optimal average expected reward per unit time and $W(\bar{n})$ is the relative value function in state $\bar{n}$ w.r.t to any fixed state $\bar{n}'$. The first term of Eq. (2.2) represents the reward accumulated while the system is in state $\bar{n}$. The second summation represents the expected reward when a patient in health-state $i \in S$ arrives into the system. The third summation gives the expected reward when a patient transitions from health-state i to j while the fourth term gives the expected reward when a

patient transitions from health-state i to either routine discharge or discharge to hospice. The last term follows after the uniformization of the CTMC into an embedded DTMC and refers to the expected reward from the system staying in the same state $\bar{n}$. Additionally, the uniformization rate ν equals $\sum_{i\in S}\lambda_i + \max_{c^f}(\sum_{(i,j)\in E} n_i r_{ij}(c_i^f)) + \max_{c^f}(\sum_{(i,j)\in \hat{E}} n_i r_{ij}(c_i^f))$. The optimal dynamic policy is in the form of a tuple $(\bar{n}, \bar{c^f})$ that denotes how many floating resources need to be allocated to each of the health states depending on the number of patients in each of the health states.

2.3.3 Modeling Assumptions

We note here some assumptions made in the formulation described in this section that allow for model tractability.

- We do not consider the time it takes for a patient to switch between wards and assume that this switching time is negligible. We believe this assumption is reasonable because time to transport patients is much smaller that a patient's length of stay.
- We turn away a patient arriving into the system with all wards full. While we can modify the model to have a 'waiting room' for the patients by creating a 'dummy unit', we would need to have a different waiting room for each patient severity type. We believe this assumption is reasonable because hospitals typically admit patients from another department, such as the Emergency Department, where patients wait if there is no room for them at the hospital.
- Finally, we assume that the patient inter-arrival time is exponentially distributed. Further, we assume that a patient's length of stay (for a fixed value of nurse-patient ratio) is exponentially distributed.
- We assume that all floating nurses are cross-trained to be able to serve across multiple different units. In reality there is often a cost associated with cross-training. We ignore this cost when evaluating the performance of the optimal dynamic policy in Section 2.4.

2.4 Numerical Analysis Of Optimal Policy

We solved the dynamic programming problem using the value iteration algorithm implemented in the Python v2.7 programming language. The convergence criterion was set as $max(W(\bar{n})_{l+1} - W(\bar{n})_l) -$

Table 2.1 Model parameters involved in performing the numerical analyses discussed in Section 2.4.

Variable Name	Variable Description	Value
p_1, p_2	Proportion of incoming patients of severity types 1, 2 entering wards 1, 2.	1/6, 5/6
$U(\bar{n})$	Health state utility accumulated while the system is in state (n_1, n_2)	$8n_1 + 2n_2$
$b_{1h}^{max}, b_{1h}^{min}$	Baseline value representing the average amount of time a patient spends in health state 1 before recovering under maximum and minimum possible values for nurse to patient ratio respectively.	40, 60
$b_{2d}^{max}, b_{2d}^{min}$	Baseline value representing the average amount of time a patient spends in health state 2 before being discharged to hospice under maximum and minimum possible values for nurse to patient ratio respectively.	30, 15
$b_{12}^{max}, b_{12}^{min}$	Baseline value representing the average amount of time a patient spends in health state 1 before moving to health state 2 under maximum and minimum possible values for nurse to patient ratio respectively.	50, 25
$b_{21}^{max}, b_{21}^{min}$	Baseline value representing theaverage amount of time a patient spends in health state 1 before moving to health state 2 under maximum and minimum possible values for nurse to patient ratio respectively.	40, 60
U'_{1h}, U'_{2d}	Fixed reward associated with the recovery and hospice discharge of an individual in ward 1 and 2 respectively.	1000, -1000

$\min(W(\bar{n})_{l+1} - W(\bar{n})_l) \leq \varepsilon$, where $W(\bar{n})_l$ refers to the value function in the l^{th} iteration of the value iteration algorithm. In addition to presenting the structure of the optimal policy, we conduct sensitivity analyses of model parameters to understand the extent of possible variation to the optimal policy. Finally, we consider the practicality of implementing the optimal dynamic policy by presenting heuristics and comparing their performance.

2.4.1 Numerical Analyses: Model Parameter Definitions

We examine the output of the model by looking at a system with 2 wards (thus 2 patient severity types) and 5 floating nurses that require allocation. We set the number of wards to 2 because it allows us to better visualize and study the structure of the optimal policy. Each ward has 4 baseline nurses that do not float and can accommodate a maximum of 10 patients. The values used for model parameters for all the experiments in Section 2.4 are outlined in Table 2.1. For some of the experiments we perform, we would like to test the impact that utilization has. However, the utilization depends on the policy. Thus, to be

able to compare policies we define a surrogate measure of system utilization $\hat{\rho}$, as

$$\hat{\rho} = \frac{\Lambda}{M \max_c (r_{1h}(M,c)) + M \max_c (r_{2d}(M,c))} \tag{2.3}$$

The numerator represents the overall arrival rate into the system and the denominator represents the rate out of the system (from ward 1 to routine discharge and from ward 2 to discharge to hospice) when there are M (the maximum allowable) patients in each ward.

2.4.2 Structure of Optimal Dynamic Policy

We examine the structure of the optimal dynamic policy (ODP) by determining how 5 floating resources should be assigned to the two wards, depending on the number of patients in each of the wards. Here we consider the case where the time spent by a patient in service is concave in the nurse-patient ratio of the ward. Recall from Section 3.1 the implication of a concave average patient service time is that nurse productivity decreases at higher workload levels.

Fig. 2.4 shows the ODP for the parameter values shown in Table 2.1 when the system experiences $\hat{\rho}$ of 80%. The color in each cell shows the number of resources that must be assigned to ward 1 depending on the number of patients present in ward 1 (y-axis) and ward 2 (x-axis). Thus, the number of resources that must be assigned to ward 2 equals 5 minus the value shown in each cell. Our first observation from Fig. 2.4 is that the policy (under these set of problem parameters) appears to follow a monotone structure in each direction. For instance, fixing the number of patients in ward 1 (y-axis) and increasing the number of patients in ward 2 (y-axis) leads to a monotone increase in number of floating resources that must be assigned to ward 2.

2.4.3 Senstivity Analysis

The structure of the optimal policy shown in Fig. 2.4 is obtained using model parameters in Table 2.1 and may not necessarily hold true under different sets of model parameter assumptions. In this section we undertake a series of sensitivity analyses to obtain a better understanding of how the structure of the optimal policy behaves under different sets of parameter values and model assumptions. The parameters we focus on for the sensitivity analyses include baseline average patient time in system, health state utility (HSU) estimates, and functional form of average patient transition time. Adjusting these parameters

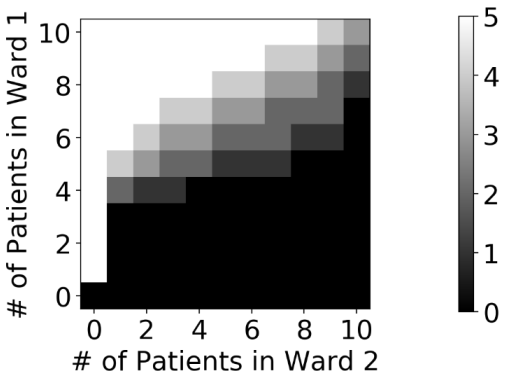


Figure 2.4 Optimal Dynamic Policy with problem parameters as outlined in Table 2.1. The color in each cell represents the number of floating resources that must be assigned to ward 1 depending on the number of patients in ward 1 (y-axis) and ward 2 (x-axis) respectively.

leads to creating unique system settings such as varying overall patient severity levels, and attempting to balance the trade-offs between having patients stay in the system versus leave the system. Additionally, we test the sensitivity of some of these parameters to varying overall patient arrival rates (thereby varying $\hat{\rho}$). A description of the experiments we performed is outlined below.

1. **Comparing the effect of time until patient improvement**

   Let R_1 be a multiplier to model parameters in Table 2.1 corresponding to the amount of time a patient spends in a state before improving in health condition. Thus, R_1 is multiplied to each of $b_{1h}^{max}, b_{1h}^{min}, b_{21}^{max}, b_{21}^{min}$. Increasing the value of R_1 means that the average time for patients to get better is longer.

2. **Comparing the effect of value of patients staying in system vs leaving**

   HSU estimates are typically the most uncertain data inputs in cost-utility models and depend on a variety of factors that must be inferred from data and/or clinical trials [Wol16]. To test the sensitivity of the HSU function, we define R_2 as a multiplier for $U(\bar{n})$. The health state utility accumulated while the system is in state (n_1, n_2) now equals $R_2(8n_1 + 2n_2)$. Thus increasing the value of R_2 leads to the model placing increased value to patients staying in system. In other words, as R_2 increases the value of keeping patients alive increases, regardless of health state, as opposed to them leaving the system to be discharged to routine or hospice.

3. **Comparing the effect of the relative health state utility values**

   Here we vary the ratio of the HSU function coefficient for the less severe ward vs that of the more

severe ward. We define R_3 as a multiplier for u_1. The health state utility accumulated while the system is in state (n_1, n_2) now equals $(R_3 * 8n_1 + 2n_2)$; thus, increasing the value of R_3 leads to the model giving higher value to patients staying in the less severe state compared to the more severe state.

4. **Comparing the effect of the functional forms for patient transition time**

   All of the optimal policy structures in experiments 1-3 consider a concave functional form for all average patient transition times (both improving and deteriorating). In this experiment we observe the structure of the optimal dynamic policy under varying levels of $\hat{\rho}$ for each of the three different functional forms for average patient transition time (concave, linear, and convex). Additionally, we consider three sub-cases of this scenario. The first is the default model with objective function outlined so far in this document. The second is when the HSU function is ignored and only one-time cost of discharge to hospice is considered.The third is when the HSU function is ignored and only the one-time reward from patient recovery is considered.

2.4.3.1 Comparing the effect of time until patient improvement.

We see from Fig. 2.5 that decreasing the value of R_1 (meaning patients transition more quickly between health states) leads to the optimal policy assigning more floating nurses to ward 2 (recall that the color white in the figure represents all float nurses being assigned to ward 1 and the color black represents the opposite). Because patient condition is improving so rapidly (as R_1 decreases) the model is able to accumulate the one-time reward of patient recovery without needing the intervention of the floating nurses. Instead (as seen in the extreme case of $R_1 = 1/100$), all the floating nurses are placed in the more severe ward to mitigate the cost of possible patient discharge to hospice.

2.4.3.2 Comparing the effect of the value of patients staying in system vs leaving.

Increasing the value of R_2 means that the model values holding patients in their respective health states more compared to them leaving the system. However, since the one-time cost of discharge to hospice is a negative value, there is still a penalty to holding patients in the more severe health state for too long as a patient may end up being discharged to hospice instead of moving to the less severe ward. As a result we see in Fig. 2.6 that in order to maximize the overall objective value the model moves all floating nurses to

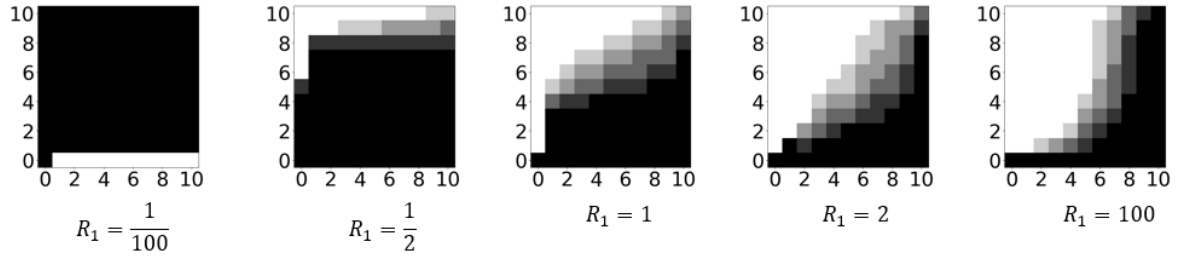


Figure 2.5 Variation in structure of ODP when problem parameters outlined in Table 2.1 are modified to change (by multiplier R_1) average amount of time it takes a patient to get better. The axes represent the number of patients in wards 1 (x-axis) and 2 (y-axis) while the colorbar represents the number of floating nurses that must be assigned to ward 1.

the more severe ward to try and reduce the number of patients being discharged to hospice.

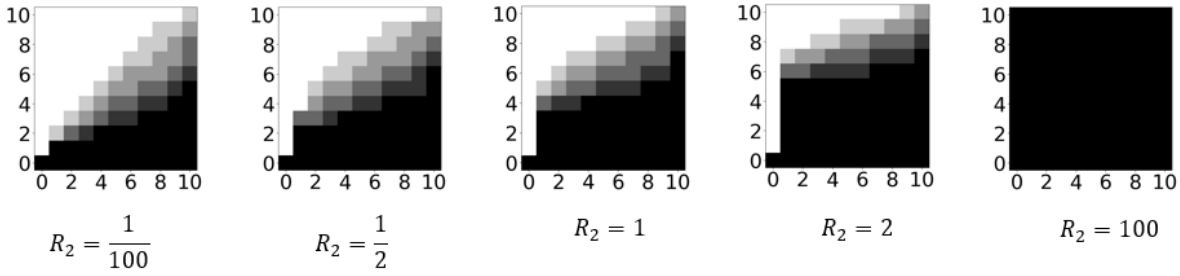


Figure 2.6 Variation in structure of ODP when problem parameters outlined in Table 2.1 are modified to increase (by multiplier R_2) the the coefficients for the Health State Utility function. The axes represent the number of patients in wards 1 (x-axis) and 2 (y-axis) while the colorbar represents the number of floating nurses that must be assigned to ward 1.

2.4.3.3 Comparing the effect of the relative health state utility values.

Increasing the value of R_3 means that the model places increased value in the less severe state relative to the more severe state. The one-time cost of patient discharge to hospice being a negative value means that model still moves all floating nurses to the more severe ward to try and reduce the number of patients being discharged to hospice. The structure of the variation in optimal policy (shown in Fig. A.1) under this experiment is similar to that seen in Fig. 2.6.

2.4.3.4 Comparing the effect of the functional forms for patient transition time.

The model outlined in this paper can be modified to account for different objectives as desired by the hospital administrators. In this section , we consider three different objectives for the MDP. The first is the default objective considered (shown in Section 2.4.2 so far with both HSU function for holding patients in units and one-time cost/reward for patient discharges. The second and third objectives ignore the HSU function and only consider patient discharge to hospice and patient discharge hto ome respectively. Within each of these objectives, we vary the functional form for a patient's average transition time and the system utilization (by changing the value of Λ).

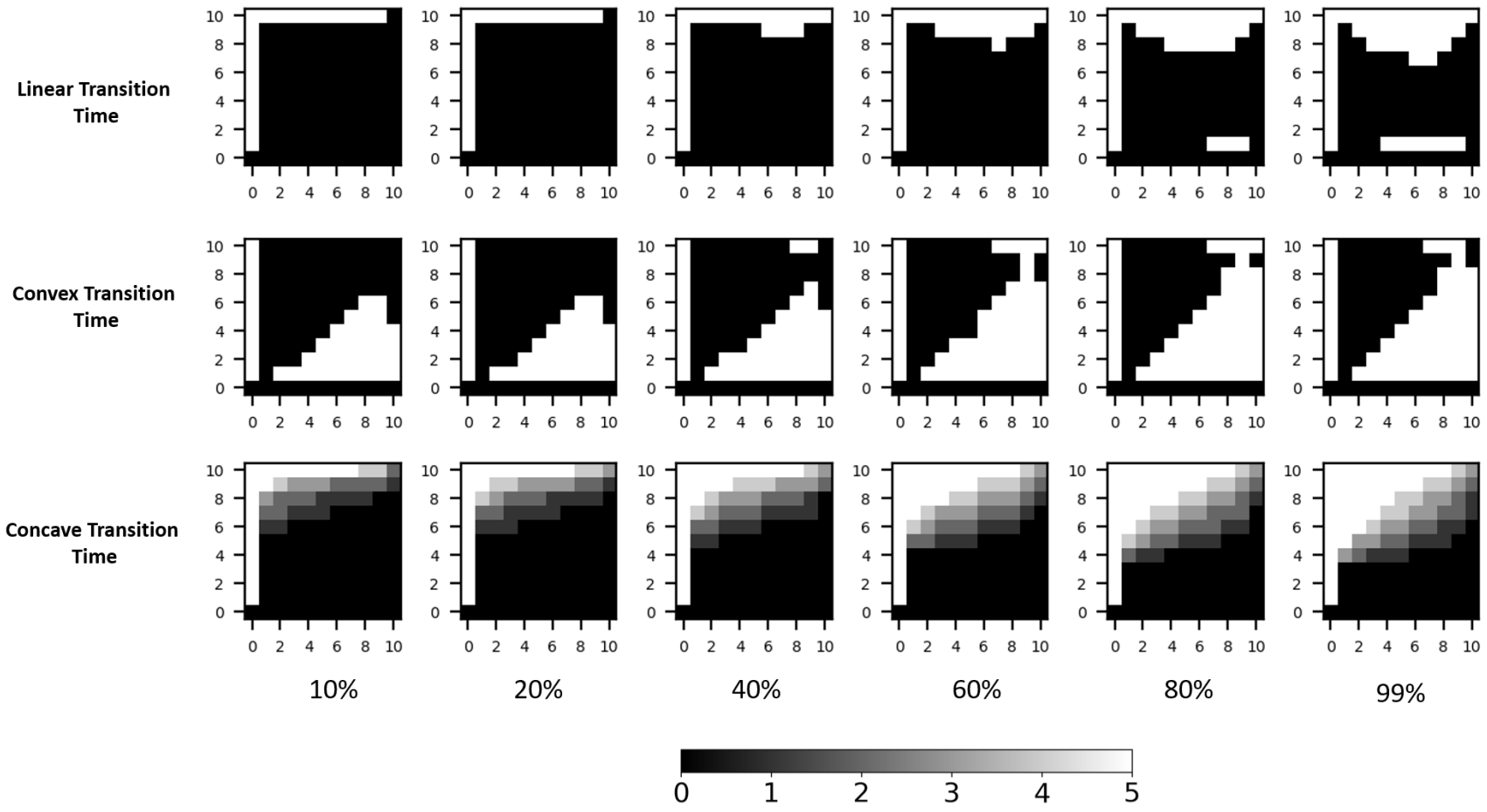


Figure 2.7 ODP for a default objective function with both holding rewards and one-time cost/reward under increasing values of system utilization $\hat{\rho}$. The colorbar shows the number of floating resources being assigned to the less severe ward. The number of patients in the less severe ward is shown on the horizontal axis while the number of patients in the more severe ward is shown on the vertical axes.

We begin by looking at Fig. 2.7 to study the role played by the functional form for average patient transition time under the default objective from Section 2.4.2. Two noteworthy observations can be made here regarding differences in structure of the optimal policy. First, we see that only in the case of concave transition time does the optimal structure appear to have situations in which the floating nurses are split between the two wards. In both the linear and convex cases, all of the floating nurses are either moved to one ward or the other. That is, in those cases it is optimal to keep floating resources together. Secondly, the structure of the optimal policy for the convex and concave cases appears to behave similarly, and sometimes have opposite recommendations to each other. To understand this effect, consider the 99% $r\hat{h}o$ case for the concave transition function in Fig. 2.7. Specifically, consider the case when there are 10 patients in the more severe ward (shown on the horizontal axis). At first (when there are a smaller number of patients in the less severe ward), the optimal policy dictates that all floating resources be assigned to the more severe ward. As the number of patients in the less severe ward (shown on the vertical axis) starts increasing, the optimal policy dictates that the floating resources start to be moved to the less severe ward. In the case of the convex function, however, the optimal policy dictates that all of the floating resources be present in the less severe ward at first (when there are less patients present there) and that all the floating nurses be moved to the more severe ward when the number of patients in the less severe ward becomes large. This difference in the behavior of the structure of the optimal policy may be attributed to the behavior of the convex and concave functional forms. Recall that on increasing the number of patients, the convex functional form means that nurse productivity decreases. This leads to the optimal policy in the convex case moving all the floating nurses over to the more severe ward only when the number of patients becomes high in the less severe ward (adding floating nurses to a high workload situation does not improve patient outcomes).

The two observations we made regarding the lack of situations in which floating nurses are split between wards in the linear and convex cases and the complementary behavior of the convex and concave policies appear to hold (shown in Fig. A.2 and Fig. A.3) across all three sets of model objectives. However, we see in the case of minimizing deaths as objective (Fig. A.2) that increasing the utilization levels leads to the optimal policy assigning all the floating nurses to the more severe ward in most cases whereas the case of maximizing discharges home (Fig. A.3) has the opposite outcome. We may recall here that our model only allows for patient discharge to hospice from the more severe ward. As a result, for the case of minimizing hospice discharges the optimal policy tries to minimize this by placing more floating nurses

in the more severe ward as utilization increases. A complementary argument may be presented in the case of the optimal policy when attempting to maximize discharges home.

2.5 Heuristic Policies: Definition and Numerical Analysis

In Section 2.4.3, we have observed that the structure of the optimal dynamic policy is highly dependent on choice of model parameters and may prove difficult to implement under special cases. Here in Section 2.5, we test the optimal dynamic policy against three different heuristics. By analyzing the optimal dynamic policy against these heuristics, we show how much the decision maker stands to lose by implementing a practical heuristic instead of the optimal policy. Furthermore, we discuss how robust any given heuristic is under varying system structure and parameter values. The three heuristics we consider are outlined below.

1. Proportional allocation (PA): The floating resources are assigned to wards in proportion to the number of patients in the wards. For instance, if there are 12 and 8 patients in wards 1 and 2, respectively, and if there are 9 floating resources to be assigned, 6 of them would be assigned to ward 1 while the remaining 3 would be assigned to ward 2. We note here that in the case of a fractional proportion, we round it to the nearest integer. For example, 4 and 3 patients in wards 1 and 2 respectively would mean that 9 floating resources would be allocated as 5 resources to ward 1 ($(4/7) \times 5$) and 4 resources to ward 2 ($(3/7) \times 5$). In the event of a tie and odd number of floating nurses, we assign an extra resource to ward 1.

2. All-or-nothing allocation (AN): Here, all the floating resources are allocated to the ward that has a greater number of patients at any point in time. Like in the previous example if there are 12 and 8 patients in wards 1 and 2, respectively, and if there are 9 floating resources to be assigned, this policy would assign all 9 resources to ward 1.

3. Optimal static allocation (OSP): This policy is static because it is fixed and is independent of the state of the system. However, we pick the best among all possible static policies by enumerating all possible static policies and picking the one that maximizes the long-run-average reward.

In addition to the objective value we obtain from the heuristics, we evaluate the performance of the heuristics by comparing the objective value obtained from employing each heuristic policy against the

objective value obtained from employing the optimal dynamic policy. The performance measure we use is percent optimality (Gap^H) where,

$$Gap^H = \frac{g^H - g^*}{g^*} \times 100$$

Here, g^* is the long-run-average reward from employing the optimal dynamic policy, and g^H is the long-run- average reward from employing any policy H. The long-run-average reward from each heuristic is evaluated as a dynamic programming problem with a restricted action space for the policy corresponding to the heuristic under consideration.

The structure of the experiments we perform to test the heuristics is similar to the sensitivity analyses in Section 2.4.3. Each heuristic is tested against the 4 experiments given in Section 2.4.3. This allows us to understand how well each heuristic performs under different experimental setups. We show some results in this section and the rest in the Appendix.

Table 2.2 A comparison of the heuristics' performance against the ODP when problem parameters outlined in Table 2.1 are modified to change (by multiplier R_1) average amount of time it takes a patient to get better.

R_1	ODP	AN		PA		OSP	
	g^*	g^H	Gap^H	g^H	Gap^H	g^H	Gap^H
1/100	65.11	50.52	22.4%	50.88	21.9%	64.1	1.6%
1/2	274.45	268.63	2.1%	270.08	1.6%	272.17	0.8%
1	181.62	181.15	0.25%	180.62	0.55%	180.72	0.49%
2	53.35	52.85	0.9%	51.29	3.9%	52.33	1.9%
100	-148.37	-151.9	2.4%	-153.77	3.6%	-153.58	3.5%

Table 2.2 compares the performance of the heuristics when increasing multiplier R_1 according to experiment 1 (comparing the effect of time until patient improvement) in Section 2.4.3. When interpreting the numbers in Table 2.2 it is helpful to consider the results from Fig. 2.5 for, in which we varied the multiplier R_1 for the optimal policy. We saw (in Fig. 2.5) in the case of $R_1 = 1/100$ that all resources were assigned to the more severe ward. In other words, the policy was optimal and appeared static. As a result, the optimality gap for OSP in the case where $R_1 = 1/100$ is much lower than the optimality gap for AN and PA. Similarly we saw that increasing the value of R_1 led to the optimal policy behaving similar to the AN policy. Table 2.2 confirms this as AN appears to perform better than OSP and PA for

increasing values of R_1. For experiments 2 and 3 (comparing the effect of the value of patients staying in system vs leaving and comparing the effect of the relative health state utility values) in Section 2.4.3, we note that the optimal policy behaves like the OSP at lower values of R_2, R_3 and behaves like AN at higher values of R_2, R_3. (Compare structure of optimal policy in Fig. 2.6 and Fig. A.1 against performance of heuristics in Table A.1.1 and Table A.1.2 for experiments 2 and 3 respectively)

Table 2.3 A comparison of the performance of heuristics against the optimal policy when the objective function includes all costs/reward as outlined in Table 2.1.

	Linear				**Concave**				**Convex**			
	ODP	**AN**	**PA**	**OSP**	**ODP**	**AN**	**PA**	**OSP**	**ODP**	**AN**	**PA**	**OSP**
10%	45.21	42.49 (6%)	42.94 (5%)	43.05 (4.8%)	41.34	37.39 (9.6%)	37.26 (9.9%)	38.94 (5.8%)	48.75	46.93 (3.7%)	47.56 (2.4%)	47.53 (2.5%)
20%	76.68	71.71 (6.5%)	72.39 (5.6%)	75.04 (2.1%)	68.40	60.97 (10.9%)	61.16 (10.6%)	67.13 (1.9%)	87.59	84.17 (3.9%)	85.23 (2.7%)	86.06 (1.7%)
40%	122.16	117.93 (3.5%)	118.48 (3%)	121.46 (0.6%)	111.14	105.25 (5.3%)	105.45 (5.1%)	110.90 (0.2%)	144.52	141.67 (2.0%)	142.79 (1.7%)	142.70 (1.3%)
60%	144.52	143.01 (1%)	143.04 (1%)	143.43 (0.7%)	134.60	131.76 (2.1%)	131.80 (2.1%)	134.52 (0.1%)	172.40	171.34 (0.6%)	171.62 (0.4%)	170.91 (0.9%)
80%	152.82	152.45 (0.2%)	152.09 (0.5%)	152.36 (0.3%)	144.09	142.71 (1.0%)	142.68 (1.0%)	144.04 (0.0%)	181.66	181.20 (0.3%)	180.66 (0.6%)	180.77 (0.5%)
99%	156.61	156.51 (0.1%)	155.96 (0.4%)	156.45 (0.1%)	148.71	147.93 (0.5%)	147.86 (0.6%)	148.66 (0.0%)	184.47	184.09 (0.2%)	183.15 (0.7%)	183.72 (0.4%)

Next we look at Table 2.3 to test the performance of heuristics against the experiment 4 that compares the effects of different functional forms for the average patient transition time. Table 2.3 specifically considers the case of the default objective that considered HSU accumulated while patients exist in the system and the one-time cost/reward of patients leaving the system. This table provides us with a holisitic view of model performance when tested against each of the three average patient transition time functional forms and varying levels of system utilization. Our first observation is that OSP appears to outperform AN and PA in 11 out of the 18 cases considered (3 functional forms across 6 system utilization levels). However, we recall here that computing the optimal static policy involves knowledge about the system such as arrival rates and functional form for average patient transition time. In contrast AN and PA are simpler from an implementation perspective and also involve less computational effort. On closer observation, we note that despite OSP outperforming AN and PA in 11 cases in Table 2.3, the percentage point difference between the optimality gap for OSP and the lower of AN and PA is greater than 3% in only 4 out of the 11 cases. Furthermore, these cases where the optimality gap is greater than

3% occurs at low utilization levels where the absolute value of system reward is lower. This indicates that the benefits gained from ease of implementation of AN and PA outweighs the benefits of lower optimality gap from OSP. The average optimality gap across all 18 cases is 3.18% for AN, 1.81% for PA and 1.32% for OSP. While OSP outperforms AN and PA on average across all cases, the margins by which it performs better is larger at lower system utilization levels. This trend is consistent even when considering the two modified objectives of of maximizing discharges home (Table A.1.3) and and minimizing discharges to hospice (Table A.1.4).

2.6 Concluding Remarks

In this paper we formulated a dynamic programming model for a hospital system with patients of different severity levels. We assign these patients on arrival to wards based on their severity. During their time in the ward, the condition of the patients may improve or deteriorate, based on which they are re-assigned units (and as a result, severity type). The patients may also leave the system, either as routine discharge (recovery) or as discharge to hospice (death). To the best of our knowledge, this work is one of the first studies that looks at incorporating nurse-patient ratios to determine patient LOS, floating nurse pools, and pooled service withing a hospital unit. We formulated a model to allocate floating resources in such a setting and tested the optimal dynamic policy against various heuristics that are easier to implement on account of practicality.

Our numerical studies show that the structure of the optimal policy varies significantly depending on input parameters. However, there appears to be little impact on the model's objective value due to implementing heuristics in place of the optimal policy. As such, using simpler heuristics, such as assigning the floating resources in proportion to the number of patients in each unit, appear to perform well. However, there is a need to validate the heuristics against real data to better capture how much the decision maker stands to lose by implementing a heuristic in place of the optimal policy. In other words, it is critical to understand the consequences of implementing a heuristic and deviating from the optimal policy. Noting that the performance on implementing a heuristic is 5.3% (as in the case of AN with 40% system utilization in Table 2.3) worse than on implementing the optimal policy must be accompanied with a discussion of what 5.3% means within the context of the system and patient outcomes. An important avenue for future research lies in validating the various model parameters shown in Table 2.1 against real

data to accurately represent the value of the optimality gap we defined when describing the heuristics in Section 2.5.

Finally, we note that we have assumed arbitrary functional forms for a patient's transition rates. There may exist other functional forms (like the sigmoid function) that better capture how the productivity of nurses in a unit changes depending on the workload levels of that unit. While it is important to consider how nurses and other hospital staff perform under varying workload levels, it is also important to be able to accurately represent the functional form of this behaviour when developing analytical models such as ours. Our goal in developing this paper is to highlight the importance of considering workload and assumptions regarding the impact of workload on patient outcomes such as time in system as this leads to significant variation in the structure of optimal policies.

CHAPTER

3

PATIENT ROUTING WITH NURSE WORKLOAD CONSIDERATIONS

3.1 Introduction

The emergency department (ED) is arguably the most operationally complex clinical setting of the modern hospital. EDs in most hospitals around the world suffer from common issues including long waits, inefficient processes and poor patient satisfaction. Increasing ED volumes has led to increasing wait times which is a sign of ED overcrowding. According to the Agency for Healthcare Research and Quality [AHR18], 90% of EDs in the country reported that they were 'holding' admitted patients in the ED while awaiting inpatient beds. This backlog of patients having to wait within the ED disrupts the efficient flow of patients throughout the hospital system. Efficient patient flow has been shown to be an important factor contributing to patient safety [Car09]. Some indicators of effective patient flow include high patient throughput, and low patient waiting times while maintaining adequate staff utilization rates and low physician idle times [Jun99].

Improving patient flow is difficult because patient arrivals into a hospital are uncertain both in timing and volume [Den13]. Despite this uncertainty, EDs have it in their power to manage the flow of patients once they arrive in order to provide effective care. Emergency departments typically stratify incoming patients into groups based on their severity. Examples of triage systems being used by hospital systems today to assess the severity of incoming patients' conditions include the Australasian Triage Scale (ATS) [Con04b], the Canadian Triage and Acuity Scale (CTAS) [JM03], the Manchester Triage System (MTS) [Par14], and the Emergency Severity Index (ESI) [Tan04]. Hospitals use such groupings of patients to route them to appropriate units within the ED for treatment. This routing (also known as 'streaming') of patients plays a vital role in improving the efficiency of an ED's operations.

While there exists extensive literature on the operational and monetary benefits of efficient patient flow and routing [Arm15; Car15; Har04a], desire for a better understanding of the impact on workload experienced by nurses and providers that results the from flow redesign is relatively new and gaining attention [Nic18]. The workload experienced by clinicians and nurses is a critical factor in the evaluation of operational metrics (e.g. clinician performance and staffing decisions) in healthcare systems [Maz16; Upe07]. High workload is associated with nurse turnover and shortages, clinician burnout, and undesired patient outcomes. Some examples of negative patient outcomes as a result of high workload include increased mortality in the ICU and during post-operative recovery, prolonged length of stay (LOS) and higher rates for procedure related infections [Hol11; LF11; Bal18; Mag17]. Quantifying workload in a healthcare setting allows us to measure and compute it as a quality of care and performance metric, and assist in the improvement of patient safety, clinician satisfaction, and performance. A critical component of studies that use systems engineering/operations research methods to quantify workload in order to describe and improve patient flow is the definition of workload. Although previous research has demonstrated ways to capture subjective workload (like NASA-TLX [Cao09]), recent literature has sought to objectively quantify workload using data available in electronic records to support real-time clinical and operational decision making in the healthcare setting [Fis19]. Additionally, since workload plays a significant role in affecting the efficiency and quality of care, there is a need to redesign routing protocols and restructure resource allocation policies while considering workload.

Our work considers an ED setting in which arriving patients need to be assigned to one of multiple different units. While the model may be generalized to account for patients having different severity levels, each of which may be routed to different sets of units, in this chapter we focus on patients of

a single (or similar) severity type(s). Such a setting is often seen in Acuity-Adaptable Units (AAUs) [Zim12] that are capable of housing patients of several severity types. Developing a mathematical model for such a setting treats patients of different severity types similarly in making routing decisions. Each unit houses a set of nurses that experience an increase in workload as more patients enter the unit. This increase in nurse workload within a unit then leads to longer patient stays in the unit. Our goal is to develop a strategy that routes incoming patients to the units in such a way that the workload experienced by the nurses is balanced and/or kept under a predetermined threshold while also ensuring that the average length of stay of a patient is minimized.

3.2 Literature Review

We begin this section by reviewing past work related to efficient routing in both health related and general service systems. For general service systems, we focus on dispatch models or skill-based routing models where incoming agents need to be assigned to one of multiple servers or server pools. For health-related literature, we restrict our focus to the routing of patients in an emergency department. In addition, we review literature that considers fairness with regards to routing, be it patients in a healthcare setting or agents/customers in general service systems. We discuss the various metrics considered by authors who studied fair routing and then discuss hospital workload and how it has been used in mathematical models.

3.2.1 Agent Routing in General Service Systems

Call centers constitute some of the earliest examples of service systems involving incoming agents requiring assignment to one or more servers. Garnett & Mandelbaum [Gar00] show examples of a small call center that are named “I", “V", “N", “X", “W" and “M" designs. The “I" and “V" designs do not require specification of a routing policy as all arriving agents have only one server or server pools to go to. All of the other designs however require some form of skill-based routing. In its broadest form, skill-based routing involves assigning the incoming agent to the most appropriate server or server pool as opposed to just assigning the incoming agent to the next available server or pool. Some commonly studied routing policies include Fastest Server First, Slowest Server First, Random, and Longest Idle Server First. Each of these policies has its merits and setbacks. For instance, Fastest Server First has been shown to lead to the lowest time in the system under specific settings (under strategic servers and

when the policy is required to be work-conserving). However, by doing so the policy disincentives hard work by requiring the most effort from the best employees, which can hurt employee retention and job satisfaction. For a more comprehensive discussion about skill-based routing policies, we refer the reader to the review of multi-class routing literature in Section 4.2.

A separate class of models that has garnered attention in more recent literature in which multiple incoming agent classes require routing to one of several different server pools are known as $\wedge$-models ('inverted-V' models). Here, servers are usually heterogeneous with multiple different server types (each type organized within a server pool). Such examples arise in situations with different experience levels for servers, or for situations with different types of incoming agent classes.

3.2.2 Patient Routing in an ED

The concept of patient routing in an ED is better captured in literature by the concept of patient flow. Jun et al. [Jun99] in their paper state 'patient routing and flow schemes' as one of three key areas that lead to improving patient outcomes and productivity within the hospital. Various authors have proposed strategies to improve the flow of patients through an ED. Some of these strategies involve using a fast track lane to reduce waiting times of low priority patients, using a combination of a fast track lane and a 'stat' lab for processing high volume tests, use of point-of-care lab testing and placing patients in a treatment area instead of sending them back to a waiting room.

The terms 'patient streaming' and 'patient segmentation' have been used in the medical community in relation to the idea of patient routing for some time [Sag12]. Patient segmentation is the process of dividing patients and processing them through an emergency department visit differently, typically depending on their acuity [Ash17]. Welch [Wel09] notes that many emergency departments had been experimenting with models that separate patients in different streams for more efficient health care delivery. In his article, he outlines different means of segmenting patients into 'service lines' each of which operates in a clearly identified geographic zone. Most segmentation strategies are based on patient severity. Oredsson et al. [Ore11] provides evidence that segregating ED patients into different streams results in reduced waiting times in comparison to a non-streamed ED model.

3.2.3 Fairness in Routing

The topic of "fairness" in large-scale service systems has been recognized by a number of authors in the past. Traditional queuing models focus on ensuring fairness from a customer's perspective. When a customer arrives and more than one server is available, the customer naturally prefers the fastest available server. However, this leads to the faster server being overworked. Recent work has stressed the importance of ensuring some form of fairness for the servers as well. One important reason for this is that perceived injustice among employees of a service industry leads to low employee satisfaction and hampers performance [Col01; CC01].

Fairness towards servers is not easily defined and depends heavily on the nature of the service/industry under consideration. Many call centers follow a longest-idle-server-first (LISF) routing policy; that is, newly arriving calls are routed to the server that has experienced the longest idle time [Arm10]. While Fastest Server First has been shown to minimize the expected time spent by an agent in the system it has also been shown to prioritize faster servers and to not distribute idle time evenly between servers. When it comes to 'fairness' among servers, Longest Idle Time First performs better as it shares the idle time among servers in proportion to service rates.

Other allocation policies considered in recent literature that attempt to achieve fairness among servers include Randomized Most Idle [Man12b], Longest Idle Pool First [Ata11] and Longest Idle Server First [Dor13]. Tseytlin [Tse09] provides a presentation of various queuing models with heterogeneous servers and also provides a literature review on fairness in service systems from the behavioral science's point of view.

Fairness towards servers can also be found in the set of literature concerning load balancing. Load balancing is a principle in resource allocation whereby the incoming load is balanced across server locations as evenly as possible. Alanyali, Hajek, et al. [Ala98] study a system with multiple customer classes and finite locations for allocation in the context of load sharing networks. They propose a Least Load Routing (LLR) policy, that assigns a customer to the location with the least load. The LLR policy and its variants are demonstrated to lead to a fluid type limit and to be asymptotically optimal.

Finally, the idea of fair routing has also been approached by a number of authors from the field of communication networks in which the network traffic load requires an even distribution among users. We refer the reader to references provided by Afek et al. [Afe99], Bartal et al. [Bar02], and Rubenstein

et al. [Rub99] for further background in this area.

3.2.4 Workload in Healthcare Analytics

Hooey et al. [Hoo17] defines workload as “the task demand of accomplishing mission requirements for the human operator". Quantifying workload within the context of healthcare delivery, specifically nurses within an emergency department, becomes challenging due to the diverse nature of tasks involved. Furthermore, there is a need to quantify workload objectively in order to be able to aid in clinical and operational decision making within hospitals. Several studies have shown that higher nursing workload has adverse effects on patient safety [Car08; AK09; Coh99]. Aside from being correlated with sub-optimal patient care, high levels of nurse workload have also been shown to lead to reduced patient satisfaction [And98]. Much of the literature exploring how nurse workload affects patient outcomes studies how nurse staffing affects patient outcomes for which there is strong evidence of correlation [Gri16]. Quantifying nurse workload is by no means an easy task, as workload is a complex construct. While the idea of nursing workload is broadly related to the number of tasks a nurse needs to complete, a universally accepted definition for nursing workload does not exist [Car08]. Different researchers and healthcare organizations use a variety of ways to define or calculate the workload experienced by the nurses. Some use the severity level of a nurse’s patients to calculate the workload [Lan04] while others have the nurses fill out questionnaires to evaluate their workload subjectively [Oet16]. Some objective workload metrics from the literature include ‘nurses per occupied bed’ [Twi09], ‘nurse to patient ratio’ [Hur08], and ‘relative value units’ (RVUs) [Gla02].

The operations research literature that considers operationalizing workload metrics via mathematical modeling to balance/minimize nurse workload is sparse. A recent paper by Fishbein et al. [Fis19] takes the important step of reviewing objective measures of workload that can be obtained from electronic records to inform operationalization of workload measurement. Most of existing work provides models that exist within the context of an inpatient [Mil12; Ago17] or home health care setting [Sir15; Pun06], or attempt to balance and minimize workload by redesigning existing staffing methods [Wri06].

This paper aims to address an essential gap in the literature concerning the lack of consideration for nursing workload when developing patient routing models in hospitals. We develop a model to route incoming patients to units in a way that ensures that nurses within different units experience a fair

allocation of workload. Two other key features that lead our work to stand out from other such projects in the literature include the pooling of resources within an ED unit and staffing-dependent patient lengths of stay.

In Section 3.3, we provide a detailed description of our definition of nursing workload and what we consider fair allocation. Then in Section 3.4, we describe an outline of a mathematical optimization model to route incoming patients into an ED while controlling for nurse workload in the wards. We then discuss methods of tractably solving the optimization model before conducting numerical experiments by analyzing the optimal policy and two heuristics.

3.3 Model Definition

We begin this section by providing an overview of patient entry and subsequent flow within our model before abstracting it mathematically by defining its structure along with the requisite parameter and variable set. Specifically, we consider patients of a similar severity type that arrive into an ED containing a set of N different wards. On arrival the patients are routed to a ward according to a decision parameter based on the number of patients and nurses in each ward. Once a patient enters a unit, they stay in service for an amount of time that is dependent on the census and staffing levels within that unit.

Fig. 3.1 shows a high-level structure of the model. Patients of a single severity type enter the system with inter-arrival time distributed exponentially with mean $1/\lambda$. We define the state of the system using an $N+1$ dimensional vector $\bar{n} = (n_1, n_2, .., n_{N+1})$ where the first N elements represent the number of patients in one of the N wards and the $(N+1)^{th}$ element represents the number of patients waiting. We note here that a patient only waits if every single ward is full. The moment a spot opens up in one of the wards, the patient at the head of the waiting line takes that spot and goes into service. On arrival, a patient is assigned to one of the available wards. Each ward i has a preset number of nurses c_i within it to tend to the patients. The choice of ward the patient is assigned to is determined probabilistically. While details regarding implementation are not in the scope of this chapter, we may think of this probability intuitively as the proportion of patients that are assigned to a ward during a specific shift period. Once a patient has been assigned to a ward, they stay there for a period of time that is exponentially distributed with a mean value that is a function of the nurse-patient ratio of the ward that they are in. The patient is discharged and assumed to have left the system once their time in service is complete.

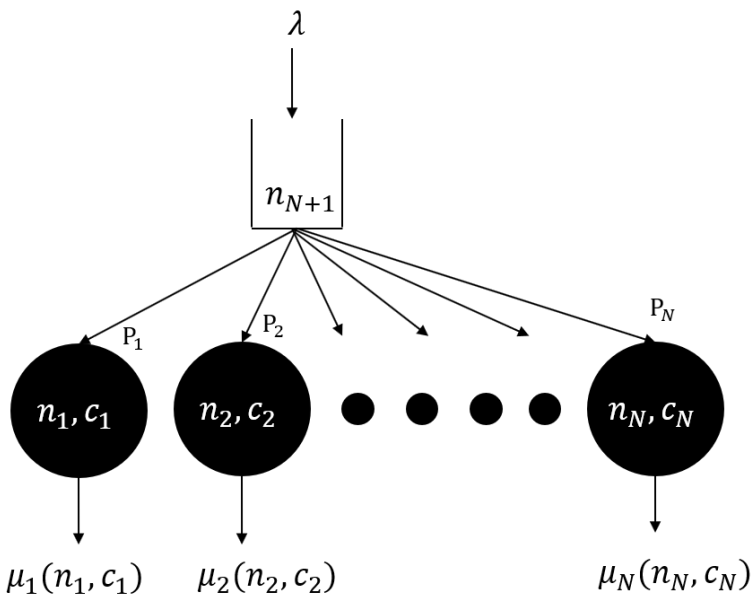


Figure 3.1 High-level model structure when we consider patients of a single severity type arriving into the system that need to be assigned to one of N different wards, each containing $n_1, n_2, ..n_N$ patients and $c_1, c_2, .., c_N$ nurses respectively.

The objective of the model we formulate in this section is to minimize the average sojourn time of patients in the system. This includes the time that a patient spends in the wait-room and the time a patient spends in service inside a ward. The model assumes that both the ward assignment probability and a patient's time in service are dependent on the workload experienced by the nurses in the wards. In the upcoming subsections we define nurse workload and how it affects the probability of patient assignment to wards and the time spent by the patients in the system. Additionally, we provide details regarding the formulation of our optimization problem.

3.3.1 Ward Assignment Probability and Patient Service Time

We begin by defining $\gamma_i(n_i, c_i)$ as the workload experienced by nurses in ward i that has a total of c_i nurses and n_i patients. We assume a general functional form for $\gamma_i(\cdot)$ with the condition that a higher nurse-patient ratio (c_i/n_i) leads to a lower nurse workload. One example is where $\gamma_i(n_i, c_i)$ is linear in the inverse of nurse-patient ratio (thus, linear in n_i/c_i). Later in Section 3.4, we assume convex and concave functions.

Next, we define the marginal increase in workload experienced by the nurses in ward i on addition of

a patient to the ward, δ_i as

$$\delta_i = \gamma_i(n_i + 1, c_i) - \gamma_i(n_i, c_i). \tag{3.1}$$

This measure is used in defining the probability of ward assignment by comparing the value of δ_i for all the wards before determining which ward a patient is to be assigned to. Specifically, if we define J as the set of wards that are not full (and thus available to accept an incoming patient), then the probability $P_i(n_1, n_2, .., n_K, \bar{\beta})$ of assigning an incoming patient to ward i is given as

$$P_i(n_1, n_2, .., n_K, \bar{\beta}) = \frac{\delta_i^{\beta_i}}{\sum_{j \in J} \delta_j^{\beta_j}}, \ -\infty \leq \beta_i \leq 0, \forall i \tag{3.2}$$

where $\bar{\beta} = (\beta_1, \beta_2, .., \beta_k)$ is a parameter that provides a measure of how to assign patients to the wards. To better understand the policy, let us consider what happens at the limiting values of β_i. First, as $\beta_i \to 0 \ \forall i$ we see that $P_i = 1/|J| \ \forall i$ and all the wards have an equal probability of having the patient assigned to them. However, as $\beta_i \to -\infty \ \forall i$, we note that P_i equals 1 if $i = \text{argmin}(\delta_j) \ \forall j$ and equals 0 otherwise. As all the values of β approach $-\infty$, the policy routes the incoming patient to the ward that is expected to experience the least marginal increase in nurse workload upon assignment of a patient. In other words, by adjusting β_i we can adjust the routing policy to result in any possible probabilistic assignment of patients to wards.

Next, we define a patient's service time. As stated earlier, the time spent by the patient in service is assumed to be exponentially distributed with a rate (μ_i) that is a function of the workload experienced by the nurses in the ward. Thus,

$$\mu_i = g(\gamma_i(n_i, c_i)).$$

While defining our model, we allow for a very general form for the function $g(\cdot)$ with the requirement that an increase in the workload experienced by the nurses $\gamma_i(n_i, c_i)$ leads to an increase in a patient's service time. We will relax this restriction later while developing the mathematical model in Chapter 4. While performing experimental analyses in Section 3.4, we consider convex and concave forms for $g(\cdot)$.

During the remainder of this chapter, we drop the reference to staffing c_i when mentioning the workload function $\gamma_i(\cdot)$ or the average patient service time function $\mu_i(\cdot)$ as staffing is assumed to be a constant in this chapter's model.

3.3.2 Continuous Time Markov Chain Formulation

As stated earlier, the objective of this work is to minimize the average time spent by a patient in the system (wait time + time in ward) while controlling for the long-run average workload experienced by the nurses in each of the wards. In this section, we model the steady-state dynamics of the system as a Markov chain. Recall that the state of the system is defined as an $N+1$-dimensional vector $\bar{n} = (n_1, n_2, .., n_{N+1})$ where the first N elements refer to the number of patients in each of the N wards and the $N+1^{th}$ element refers to the number of patients waiting.

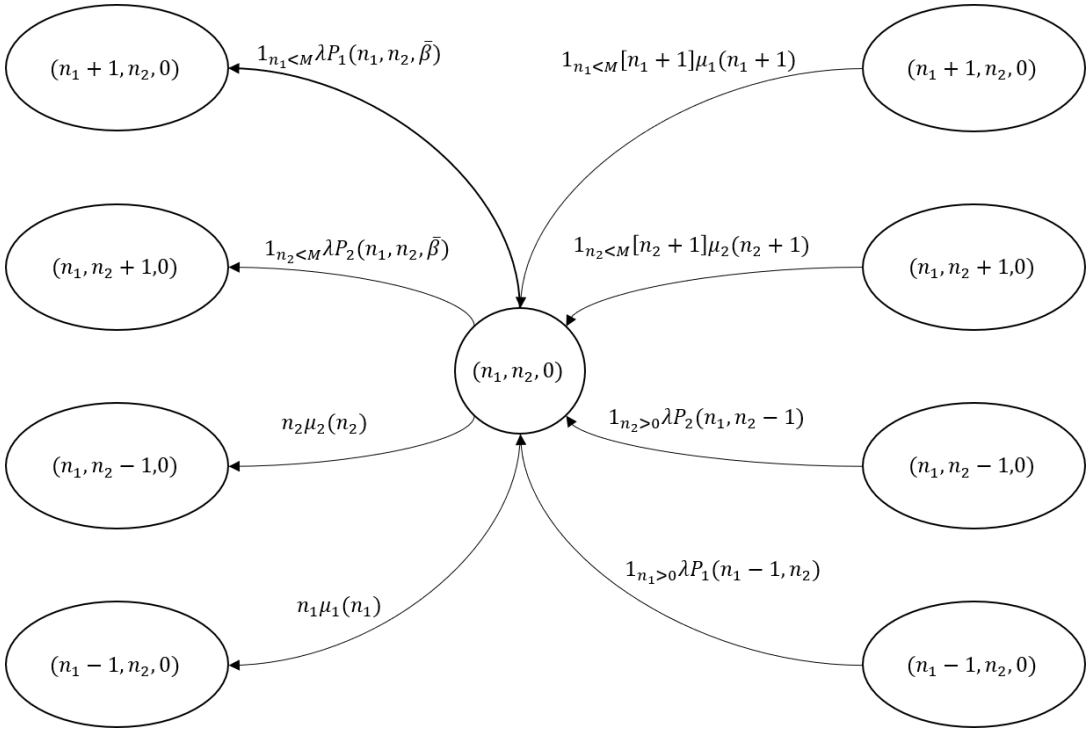


Figure 3.2 CTMC transition diagrams when at least one ward is available.

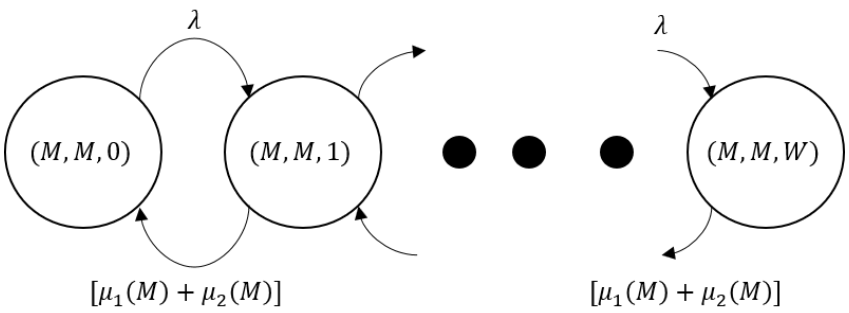


Figure 3.3 CTMC transition diagram when all wards are full and patients have to wait.

In order to simplify notation and make our understanding of the model easier, let us start by considering a two ward system, with each having a maximum capacity of M patients per ward and a total of W waiting spots. The state of the system is now the set $S = \{(n_1, n_2, n_3) : n_1 \leq M, n_2 \leq M, n_3 = 0\} \cup \{(n_1, n_2, n_3) : n_1 = M, n_2 = M, 0 < n_3 \leq W\}$. Let $\bar{\alpha} = \{\alpha_s | s \in S\}$ be the steady state probability of the system finding itself in state $s \in S$ and $Q(\bar{\beta})$ be the rate transition matrix associated with the possible state transitions of the system, for a given routing policy $\bar{\beta}$ as shown in Fig. 3.2 and Fig. 3.3. Now, $\bar{\alpha}$ may be obtained by solving the system of equations $\bar{\alpha}Q(\bar{\alpha}) = 0;\ \bar{\alpha}\bar{e} = 1$.

Given the steady state distribution of the system being in state $s \in S$, we have the long-run average number of patients in the system to be $\sum_{i,j,k}[i+j+k]\alpha_{i,j,k}$. With the blocking probability (on account of all wards and wait room being full) equaling $\alpha_{M,M,W}$, little's law gives us the long-run average time spent by a patient in the system to be $E[T] = \frac{\sum_{i,j,k}[i+j+k]\alpha_{i,j,k}}{\lambda[1-\alpha_{M,M,W}]}$. This is the objective that we wish to minimize. Finally, we are able to compute desired workload metrics such as long-run average workload in each ward and difference in workload experienced by nurses of any two wards. The long-run average workload experienced by nurses in wards 1 and 2 (denoted by $E[\Gamma_1]$, $E[\Gamma_2]$) are given as $\sum_i \sum_j \sum_k \alpha_{i,j,k}\gamma_1(i)$ and $\sum_i \sum_j \sum_k \alpha_{i,j,k}\gamma_2(i)$. Furthermore, the absolute difference in workload experienced by nurses of two difference wards may be computed as $\sum_i \sum_j \sum_k \alpha_{i,j,k}|\gamma_1(i) - \gamma_1(j)|$. We note here that our model offers flexibility in choice of workload constraint.

We are now able to define our optimization model which is provided below.

$$
\begin{aligned}
\underset{\bar{\alpha},\bar{\beta}}{\text{minimize}} \quad & E[T] = \frac{\sum_{i,j,k}[i+j+k]\alpha_{i,j,k}}{\lambda[1-\alpha_{M,M,W}]} & \\
\text{subject to} \quad & \bar{\alpha}Q(\bar{\beta}) = 0 & (3.3.1)\\
& \bar{\alpha}e = 1 & (3.3.2)\\
& \Psi(\Gamma_1,\Gamma_2) \in \psi, & (3.3.3)\\
& 0 \leq \alpha_{i,j,k} \leq 1, \qquad i \leq M, j \leq M, k \leq W & (3.3.4)\\
& -\infty \leq \beta_1, \beta_2 \leq 0 &
\end{aligned}
\tag{3.3}
$$

The objective function of the optimization model in Eq. (3.3) aims to minimize the long-run average time spent by a patient in the system. Constraints (3.3.1) and (3.3.2) represent the rate balance equations for the CTMC which are elaborated in optimization model in Eq. (C.1) provided in Appendix C. Constraint (3.3.3) is a general workload constraint stating that the value of a desired function $\Psi(\Gamma_1,\Gamma_2)$ of the workload of two wards be within a desired range in set ψ.

While conducting numerical experiments in this chapter, we consider Ψ and ψ such that the workload is balanced across wards. For the two ward case considered during the experimental analyses in Section 3.4, $\Psi(\Gamma_1,\Gamma_2) \in \psi$ is formulated as $\sum_i \sum_j \sum_k \alpha_{i,j,k}|\gamma_1(i) - \gamma_2(j)| \leq \gamma_d^*$. This constraint ensures that the long-run absolute average difference in workload between wards 1 and 2 be kept under a threshold γ_d^*. Another example of this constraint would be if $\Psi(\Gamma_1,\Gamma_2) \in \psi$ were formulated to $\{\sum_i \sum_j \sum_k \alpha_{i,j,k}\gamma_l(i) \leq \gamma_l^* \; \forall \, l\}$, which requires the long-run average workload experienced by nurses in ward $l \in \{1,2\}$ be kept under a pre-defined threshold constant γ_l^*.

Finally, we set upper and lower bounds for β_1, β_2 in Eq. (3.3) to be 0 and $-\infty$ by definition.

3.3.3 Stationary Distribution Using LDQBD Processes

The optimization model defined in the previous section is difficult to solve because of the large number of non-linear equality constraints. In particular, the crux of the difficulty is due to the set of constraints $\bar{\alpha}Q(\bar{\beta}) = 0; \alpha e = 1$. Computing the steady state probabilities $\bar{\alpha}$, given $\bar{\beta}$, becomes difficult when the discrete state space becomes extremely large, and often requires leveraging some form of a special problem structure (e.g, birth-death processes [Cra14]). In this subsection we are able to show that our

problem may be represented as a level dependent quasi birth-death processes (LDQBD). While a birth-death process is a univariate markov process whose transition probability matrix exhibits a tri-diagonal structure, a quasi birth-death process is the corresponding bivariate form where the probability transition matrix exhibits a block tri-diagonal structure. The LDQBD process is a generalization of the QBD process where each level (row block) requires an explicit and often unique definition. For more details, we refer the reader to an article by Kharoufeh [Kha10].

To convert our problem in the form of a LDQBD process, we begin by defining an appropriate ordering of the state space. Recall that in the case of a 2 ward example, our state space is 3 dimensional with the first dimension representing the number of patients in unit 1, the second dimension representing the number of patients in unit 2, and the third dimension representing the number of patients currently waiting for an available ward. The ordering of our state space then is done in a such a way that for a particular value of n_1, we iterate through all possible values of n_2 before listing the set of states with the next value for n_1. Finally, all of the states with a non-zero value for n_3 are listed at the end when n_1 and n_2 equal M. Thus, the ordering becomes

$$\begin{aligned}\{&(0,0,0),(0,1,0),(0,2,0),...,(0,M,0),\\ &\quad(1,0,0),(1,1,0),(1,2,0),..,(1,M,0),\\ &\qquad...,\\ &\qquad\quad(M,0,0),(M,1,0),(M,2,0),..,(M,M,0),(M,M,1),(0,M,2),...,(M,M,W)\}\end{aligned} \tag{3.4}$$

A Quasi birth-death process is defined using phases and levels such that the process can only go up or down one level at a time. Within each level is a set of phases that do not have this same restriction. In the context of our problem, we define our levels as the number of patients in the first unit and the phases within each level as the number of patients in the second unit. For a problem setting involving 2 wards with a maximum of ($M =$) 3 patients in each ward and a maximum of ($W =$) 3 waiting spots, the rate transition matrix $Q(\bar{\beta})$ thus takes the structure shown in Fig. 3.4.

We see from Fig. 3.4 that the rate transition matrix $Q(\bar{\beta})$ from Eq. (3.3) maintains a special block diagonal structure. All the non-zero cells in the matrix are shown via black squares in the figure while all the remaining cells have values of 0. As a result, we are able show that the CTMC has a special structure and can be formulated as a level dependent quasi birth death process (LDQBD). Each of the $M+1$ levels

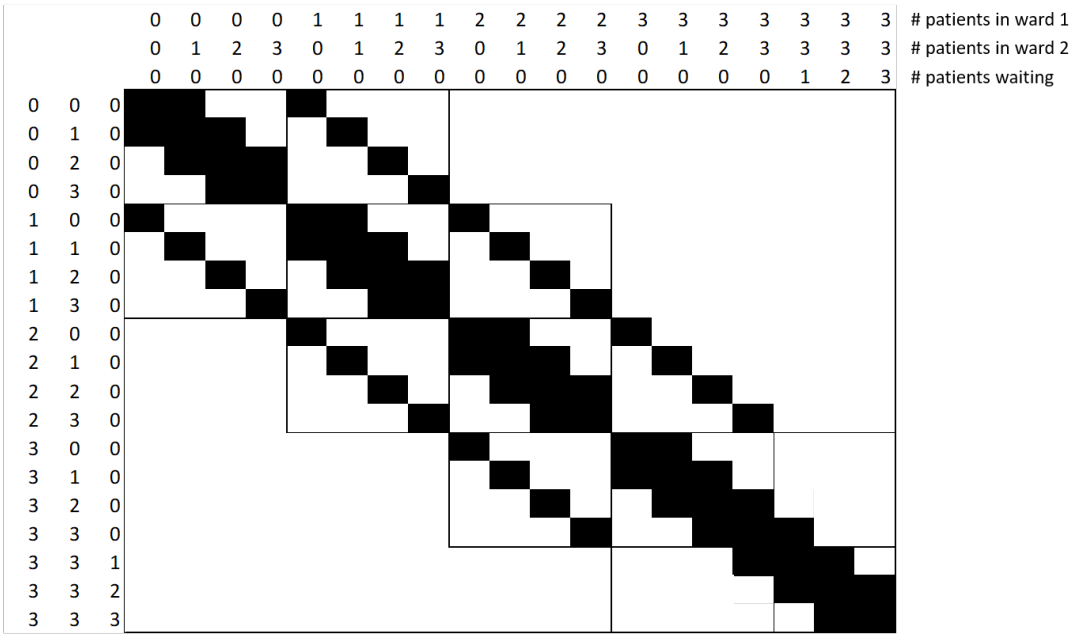


Figure 3.4 Structure of the rate transition matrix $Q(\bar{\beta})$ under a 2-ward setting wtih a maximum of 2 patients in each ward and a maximum of 3 waiting spots. A cell within the matrix $Q(\bar{\beta})$ is marked as black if it is non-empty and is marked white otherwise.

has $M+1$ phases except for the last level which has $M+1+W$ phases to represent the number of waiting patients. The matrix $Q(\bar{\beta})$ can be re-written via levels $L_0, L_1, .., L_M$ as

$$
Q(\bar{\beta}) = \begin{array}{c} \\ L_0 \\ L_1 \\ L_2 \\ \vdots \\ L_{M-1} \\ L_M \end{array}
\begin{array}{c}
\begin{array}{cccccc} L_0 & L_1 & L_2 & \dots & L_{M-1} & L_M \end{array} \\
\begin{pmatrix}
A_1(0,\bar{\beta}) & A_0(0,\bar{\beta}) & \dots & & & \\
A_2(1,\bar{\beta}) & A_1(1,\bar{\beta}) & A_0(1,\bar{\beta}) & & & \\
 & A_2(2,\bar{\beta}) & A_1(2,\bar{\beta}) & & & \vdots \\
\vdots & & & \ddots & & \\
 & & & \dots & A_1(M-1,\bar{\beta}) & A_0(M-1,\bar{\beta}) \\
 & & & \dots & A_1(M,\bar{\beta}) & A_0(M,\bar{\beta})
\end{pmatrix}
\end{array}
$$

$A_0(i,\bar{\beta})$, $A_1(i,\bar{\beta})$ and $A_2(i,\bar{\beta})$ are the generator matrices for the QBD process, where i gives the level (number of patients in the first unit). Since the generator matrices of our QBD is dependent on the level i, ours is a level-dependent quasi birth-death process. Appendix B outlines the structure for each of $A_0(i,\bar{\beta})$, $A_1(i,\bar{\beta})$ and $A_2(i,\bar{\beta})$.

3.3.4 Summary of Solution Procedure to Obtain Optimal $\bar{\beta}$

While level-independent quasi birth death processes are fairly easy to solve, on account of the existance of closed form expressions for the steady state probabilities, level-dependent quasi birth death (LDQBD) processes are a little more complex. Kharoufeh [Kha10] provides a good description of both discrete- and continuous- time LDQBD processes and algorithmic approaches and extensions of solution methodologies. When the number of levels and the number of phases are both finite, as is the case in our model, Gaver et al. [Gav84] provides a stable algorithm to compute the steady state probability for a continuous-time case. Given a $\bar{\beta}$, we use this algorithm to solve for $\bar{\alpha}$ to then compute the expected time in the system. The general outline of our methodology using Gaver's algorithm is stated as follows.

Algorithm 1 Gaver's Algorithm to solve for $\bar{\alpha}$ given $\bar{\beta}$

Input $\{A_0(i,\bar{\beta}), A_0(i,\bar{\beta}), A_0(i,\bar{\beta})\}, \forall i \in 1,..,M$

Output $\bar{\alpha}$

procedure 1

    $C_0^{-1} \leftarrow A_1^{-1}(0,\bar{\beta})$

    for $m \in \{2,..,M\}$ **do**

        $-C_{m-1}^{-1} \leftarrow -A_1^{-1}(m-1,\bar{\beta}) - A_2^{-1}(m-1,\bar{\beta})C_{m-2}^{-1}A_0^{-1}(m-2,\bar{\beta}))$

    end for

    $C_M \leftarrow A_1(M,\bar{\beta}) - A_2(M,\bar{\beta})C_{M-1}^{-1}A_0(M-1,\bar{\beta})$

    $\alpha_M \leftarrow$ solution of $\{\alpha_M C_M = 0;\ \alpha_M \mathbf{e} = 1\}$

    for $m \in \{M-1,..,0\}$ **do**

        $\alpha_m \leftarrow \pi_{m+1}A_2(M+1,\bar{\beta})(-C_m^{-1})$

        renormalize α_m;

    end for

    $\hat{\alpha} \leftarrow \sum_{m=0}^{M} \alpha_m \mathbf{e}$

    for $m \in \{0,..,M\}$ **do**

        $\alpha_m \leftarrow \hat{\alpha}\alpha_m$;

    end for

    return $\bar{\alpha}$

end procedure

The resulting vector $\bar{\alpha} = \{\alpha_m \forall m = 0,..,M\}$ corresponds to the vector of the stationary probability distribution. Algorithm 1 provides an efficient means of computing the stationary distribution $\bar{\alpha}$ given a LDQBD formulation. This means that given a $\bar{\beta}$, we are able to express our patient routing model as a LDQBD following which we can compute the stationary distribution $\bar{\alpha}$ using Algorithm 1. Knowing $\bar{\alpha}$ now allows us to compute the average patient time in system and average nurse workload in both wards using the LDQBD formulation.

Thus far we have discussed ways of computing average patient time in system and average nurse workload of both wards for a given value of $\bar{\beta}$. However, our objective is to solve the constrained

optimization model in Eq. (3.3). One way of doing this is by using black-box optimization in which the black-box computes the patient time in system using Algorithm 1 for different values over the search space of $\bar{\beta}$. However, our numerical experiments in Section 3.4 assume small problem instances that allow us to enumerate over the space of $\bar{\beta}$.

Specifically, we enumerate each $\beta_i \in \bar{\beta}$ through all values between 0 and $-\infty$ with a step size of ε. For $\bar{\beta}$, we use Algorithm 1 to calculate the average patient time in system ($E[T]$). Next we check if the workload constraints ($\Psi(\Gamma_1, \Gamma_2) \in \gamma^*$) are violated, in which case the corresponding value of $E[T]$ is marked as 'infeasible'. Finally, we look at all the 'feasible' values of $E[T]$ and pick out the lowest value as the optimal objective value and use the corresponding $\bar{\beta}$ as the optimal decision variable value. We note here that in Section 3.4, we assumed a step size $\varepsilon = 0.1$ for computational tractability and because we observed that this step-size was sufficient to provide a smooth distribution for $E[T]$ when enumerating over the space of $\bar{\beta}$. Fig. 3.5 depicts a flowchart outlining this solution procedure.

Finally, we note here that we determined the lower bound on $\bar{\beta}$ ($-\infty$) to be a value 'small enough' beyond which changing $\bar{\beta}$ leads to an objective function ($E[T]$) change of less than 0.01%. Specifically we use the value of -50 as a surrogate for $-\infty$.

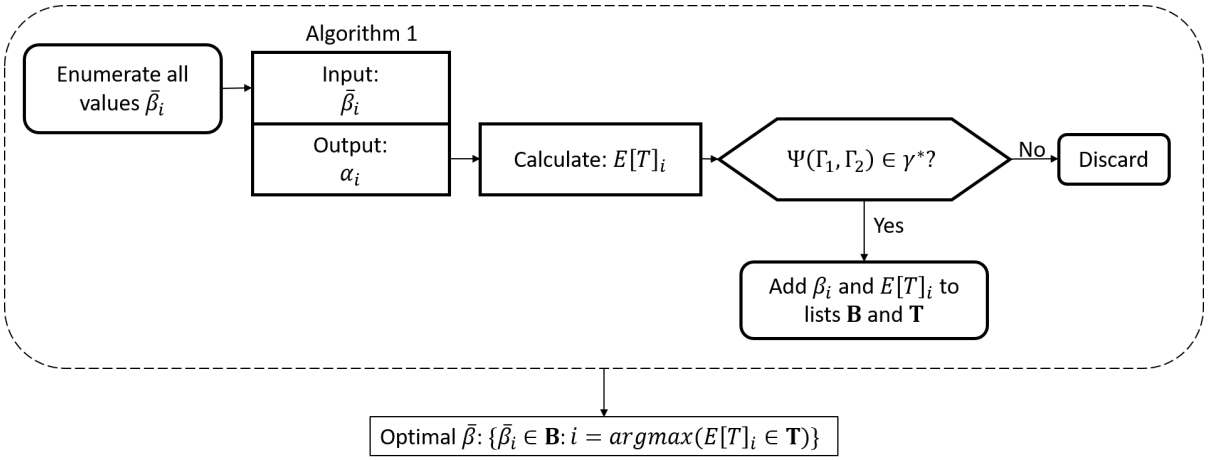


Figure 3.5 Flowchart outlining the complete solution structure to obtain optimal values of $\bar{\beta}$ utilizing Algorithm 1.

3.4 Experimental Analyses

We have two primary goals when conducting numerical experiments. The first is to compare the performance of two heuristics against the optimal policy obtained by solving the optimization model in Eq. (3.3) *without* enforcing the workload threshold constraints. This comparison allows us to test the extent to which the optimal policy can reduce the average patient sojourn time without controlling for workload. The two heuristics that we compare the optimal policy to are myopic workload routing policy (MWRP) and proportion routing policy (PRP). Recall from section 3.3.1 that while defining our patient routing policy, we specified the range for $\beta_i \in \bar{\beta}$ to be between $-\infty$ and 0. $\beta_i \to -\infty$ refers to a policy that routes incoming patients to the ward expected to have the least marginal increase in nurse workload on assigning the patient to it (MWRP). The second heuristic we consider is proportional routing policy (PRP) where we route incoming patients to wards in proportion to the number of nurses in each ward. For instance, if ward 1 has 4 nurses and ward 2 has 2 nurses and neither ward is full, an incoming patient is routed to ward 1 with probability 2/3 and is routed to ward 2 with probability 1/3.

Our second goal is to study the performance of the optimal policy when considering workload balance constraints. Specifically, we constrain the optimization model to have an absolute difference between long-run ward workloads to be under a certain balance target γ^*. In order to determine an effective value for γ^*, we use the long-run difference in workload experienced by nurses in both wards under a myopic policy that always sends an incoming patient to the ward with the lowest marginal increase in workload (MWRP). Solving the optimization model thus provides us with a routing policy that minimizes the average patient sojourn time while ensuring that the nurses in the wards experience workload levels that are more balanced than under MWRP.

We conduct all of the experiments in this section for a 2-ward setting with each ward having a maximum capacity of 15 patients and the waiting room having a maximum capacity of 10 patients. We only consider two wards here as it allows for computational tractability and ease of interpretation of the results. We experiment with different functional forms (convex and concave) for the workload function. Specifically, we assume the workload function to be convex or concave in the inverse of nurse-patient ratio as $\gamma_i(n_i) = b_i \times f(n_i/c_i)$, where $f(x)$ equals either x^2 (for a convex function) or $\sqrt{x}$ (for a concave function). Here, b_i is the baseline amount of workload experienced by a nurse in ward i when the nurse-patient ratio equals 1. We assume values of $b_1 = 10$ and $b_2 = 2$ to indicate that nurses in ward 1

have a higher baseline level of workload.

Like workload, we assume that the average amount of time a patient spends in the ED ward before departing is either a convex or concave function in the nurse-patient ratio. We assume a similar form for the rate function as in Section 2.3.1 and use the functional form corresponding to the rate at which a patient's health condition deteriorates. This is because we assume in this chapter that the longer a patient spends in the ward, the worse the patient's health condition gets. We thus define the rate $\mu(n_i)$ at which the service rate corresponding to a patient's time in system while in ward i with n_i total patients as

$$\mu(n_i) = \left(a_i^{min} + (a_i^{max} - a_i^{min}) \times q\left(\frac{n_i/c_i - \varepsilon^{min}}{\varepsilon^{max} - \varepsilon^{min}}\right)\right)^{-1}. \tag{3.5}$$

Here, a_i^{max} is a baseline value representing the amount of time it takes a patient in ward i to leave when the patient-nurse ratio (which is merely the inverse of the nurse-patient ratio) ($\frac{n_i}{c_i}$) is the maximum possible (ε^{max}) and a_i^{min} is a baseline value representing the amount of time it takes a patient to leave ward i when the patient-nurse ratio ($\frac{n_i}{c_i}$) is the minimum possible (ε^{min}). Convexity or concavity of the function $q(\cdot)$ determines how a patient's time in system changes on increasing patient-nurse ratio. $q(x) = x^2$ leads to a convex function while $q(x) = \sqrt{x}$ leads to a concave function.

Recall from Chapter 2 that the differences in functional form are representative of the differences in how nurses within a ward perform as their workload (for which patient-nurse ratio can be thought of as a proxy) increases. A convex form, for instance, means that the marginal productivity of nurses decreases at increasing workload levels. In other words, a convex form leads to a situation of increasing marginal effort (IME) as the effort required from the nurses to treat each additional patient increases. A concave form, however, means that the marginal productivity of nurses increases at higher workload levels. In other words, a concave form leads to a situation of decreasing marginal effort (DME) as the effort required from nurses to treat each additional patient decreases. Another way of looking at convexity (IME) and concavity (DME) is from the point of view of how fast patient time in service or nurse workload increases. DME means that the increase is rapid at lower occupancy levels and slows down as more patients are added. IME means that the increase is slow at lower occupancy levels and speeds up as more patients are added.

We consider four different scenarios that test the change in a patient's service time and nurse workload on increasing the number of patients (and therefore nurse-patient ratio) of the ward being considered.

The four scenarios are shown pictorially in Fig. 3.6 and are described below. We note here that though we have not shown pictorially how nurse workload changes on increasing the number of patients, the functional form is similar to the average patient sojourn time shown in Fig. 3.6.

- Scenario 1: Scenario Fig. 3.6a, which we will refer to during the rest of this chapter as 'both wards DME', considers concave functions for patient time in service and nurse workload in both wards. We assume here that at max occupancy levels, patients of ward 1 spend longer on average in service than patients of ward 2 on account of nurses in ward 1 experiencing higher workload levels than nurses in ward 2.

- Scenario 2: Scenario Fig. 3.6b, which we will refer to as 'ward 1 IME; ward 2 DME', considers a convex function for ward 1 and a concave function for ward 2. Recall that ward 1 has more nurses than ward 2.

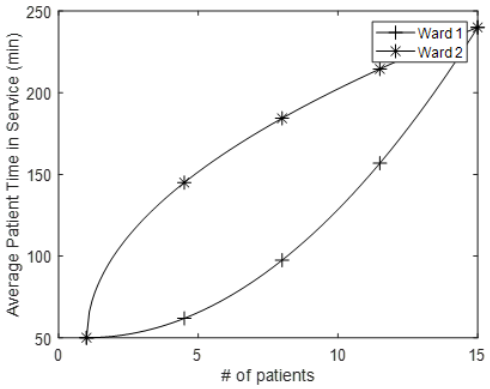
- Scenario 3: Scenario Fig. 3.6c, which we will refer to as 'ward 1 DME; ward 2 IME', is the opposite of scenario 2. Here, the ward with the more nurses (ward 1) is concave while the ward with fewer nurses (ward 2) is convex.

- Scenario 4: Scenario Fig. 3.6d, which we will refer to as 'both wards IME', is the opposite of scenario 1. Here, both wards are assumed to have convex functions with patients of ward 1 spending longer on average in the the ward at higher occupancy levels.

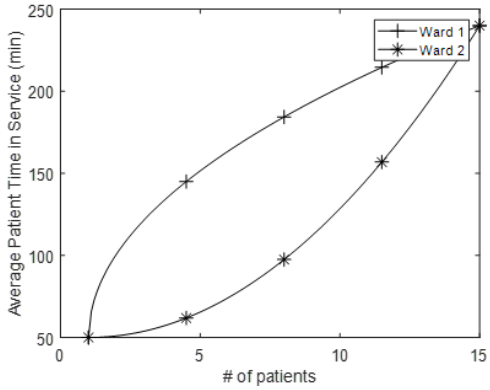
Finally, as part of our experimental design we change the rate of patients arriving into the system to study the effect of increasing the 'strain' being placed on the system. To capture this, we develop a proxy measure of system utilization as a function of the arrival rate which we define as $\hat{\rho}$. Where

$$\hat{\rho} = \frac{\lambda}{\sum_{i=1}^{K} \frac{1}{2} M \mu(\frac{1}{2}M)},$$

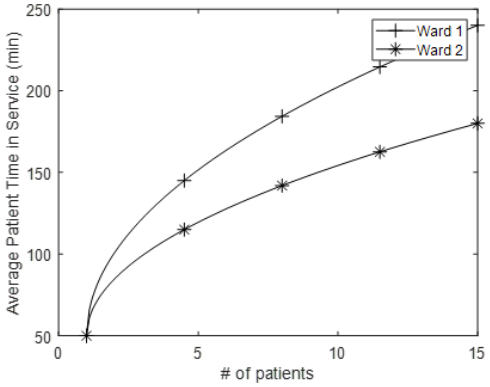
the numerator is the arrival rate into the system and the denominator is a measure of total service rate when all the wards are half-full. While $\hat{\rho}$ is merely a proxy measure, it allows us to compare the performance of our different routing policies within reasonable ranges of λ.



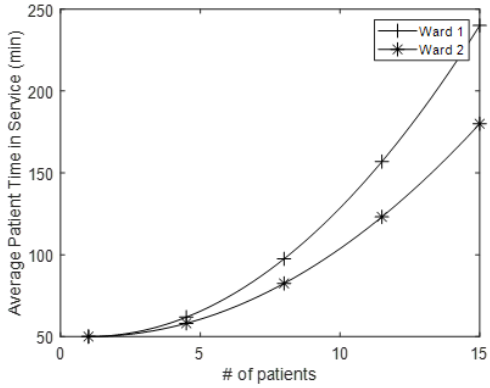
(a) Ward 1 IME; ward 2 DME



(b) Ward 1 DME; ward 2 IME



(c) Both wards DME



(d) Both wards IME

Figure 3.6 Experimental scenarios for functional form of patient time in system on increasing number of patients. We note here that all patients are assumed to have the same value for baseline average time in ward when they are the only patient in the ward.

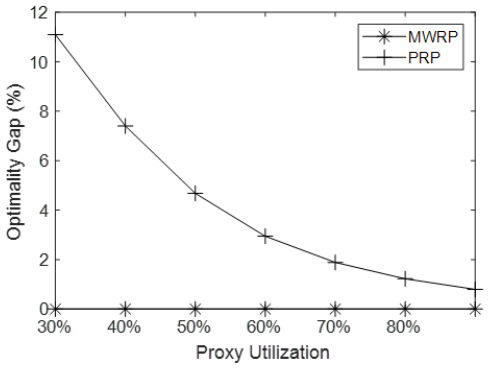
3.4.1 Experiment 1: Study of the ORP in Absence of Workload Constraints

In this experiment we study the behavior of the optimal routing policy (ORP) and compare it against two heuristics (MWRP and PRP) in the absence of workload constraints. We conduct four scenarios of patient time in system and nurse workload function which are shown in Fig. 3.6. In each case, we vary the proxy utilization of the system between 30% and 90% in increments of 10% to understand how the optimal policy behaves compared to the two heuristics on increasing the strain being placed on the system. Under each scenario we look at two performance measures - (1) optimality gap measuring the percentage increase in long-run average patient sojourn time under each heuristic when compared against ORP, and (2) the percentage increase in long-run average difference in workload under each

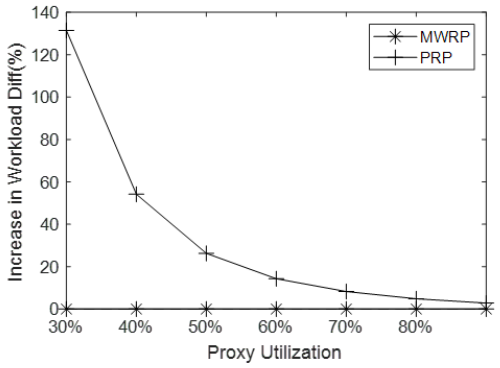
heuristic when compared against ORP. The first measure gives an idea of the performance of the optimal policy by measuring how much longer a patient spends on the system on average. The second measure gives an idea of how well the optimal policy performs in balancing workload. A positive value for percentage increase in long-run average difference in workload means that the heuristic does a poorer job of balancing workload compared to the optimal policy while the reverse is true in the event of a negative value.

3.4.1.1 Scenario 1: Ward 1 IME; ward 2 DME

Fig. 3.7 shows the optimality gap and the increase in long-run average difference in workload under each heuristic when compared against ORP. Our first observation is that at higher levels of system utilization, both heuristics behave similarly to ORP. This is because at high levels of utilization both wards are either full or close to being full and the decision of which ward to route the patients to does not affect the average patient LOS. We continue to make this observation for average patient sojourn time under all scenarios indicating that at higher system utilization levels, the choice of routing policy does not play a major role as far as average patient sojourn time is concerned.



(a) Optimality gap measuring the percentage increase in long-run average patient time in system under heuristic when compared against ORP



(b) Percentage increase in long-run average difference under each heuristic when compared against ORP

Figure 3.7 Ward 1 IME; Ward 2 DME. We see that MWRP performs very similar to ORP while PRP performs worse both in reducing patient sojourn time and in balancing nurse workload between wards.

Our second observation (not shown in figure) is that the values of β_1, β_2 under ORP equal $-50, -50$. In other words, ORP appears to behave like MWRP at all utilization levels. Fig. 3.7 confirms that MWRP

behaves similar to ORP with a value of zero (under all system utilization levels) for optimality gap and percentage increase in long-run average workload difference.

Our third observation is that the values of $P_1(n_1,n_2,\hat{\beta}), P_2(n_1,n_2,\hat{\beta})$ for all values of n_1,n_2 are such that the policy always routes patients to ward 1 and only routes patients to ward 2 if ward 1 is at maximum capacity. The reason ORP attempts to fill up ward 1 before routing patients to ward 2 can be understood as follows. The combination of DME nature and fewer nurses for ward 2 means that at lower levels of system utilization, adding patients to ward 2 leads to sharper increases in patient service time compared to adding patients to ward 1. Thus at lower levels of utilization, it appears prudent to send incoming patients to ward 1 (thus filling up the IME ward) as it has more nurses and patient wait time does not increase as rapidly on adding patients as it does in ward 2.

To mathematically understand why ORP fills up the IME ward before assigning patients to the DME ward, we consider the heuristic that ORP appears to behave like - the MWRP. To understand why MWRP fills up ward 1 before routing patients to ward 2, we begin by recalling that under MWRP, assignment of an incoming patient to a ward is made depending on which ward is expected to have the smallest marginal increase in workload on adding a patient to it. In other words for our two ward case, the ward i that an incoming patient must be sent to is determined by the relation

$$i = \underset{i\in\{1,2\}}{\operatorname{argmin}} \left(\gamma_i(n_i+1) - \gamma_i(n_i)\right). \tag{3.6}$$

Now, since ward 1 is IME (convex),

$$\gamma_1(n_1+1) - \gamma_1(n_1) = \frac{b_1}{c_1^2}\Big((n_1+1)^2 - n_1^2\Big) = \frac{b_1}{c_1^2}\Big(2n_1+1\Big). \tag{3.7}$$

Similarly, ward 2 being DME (concave) implies that

$$\gamma_2(n_2+1) - \gamma_2(n_2) = \frac{b_2}{c_2^2}\Big(\sqrt{n_2+1} - \sqrt{n_2}\Big). \tag{3.8}$$

Now, MWRP dictates sending an incoming patient to ward 1 if

$$\frac{b_1}{c_1^2}\Big(2n_1+1\Big) \leq \frac{b_2}{c_2^2}\Big(\sqrt{n_2+1} - \sqrt{n_2}\Big).$$

Rearranging the terms gives us the following relationship between n_1 and n_2 under which an incoming patient will be sent to ward 1.

$$n_1 \geq \frac{1}{2}\left(\frac{b_2 c_1^2}{b_1 c_2^2}\left(\sqrt{n_2+1}-\sqrt{n_2}\right)-1\right). \tag{3.9}$$

Equation (3.9) provides us with conditions under which an incoming patient will be sent to ward 1 once ward 2 has a certain number of nurses. Now, consider the right hand side of the equation. If the RHS is negative, then it is clear that incoming patients will be sent to ward 1 no matter how many patients ward 1 already has. In other words under the following condition, $\frac{b_2 c_1^2}{b_1 c_2^2} \leq \frac{1}{\sqrt{n_2+1}-\sqrt{n_2}}$, MWRP always dictates sending the next incoming patient to ward 1 after ward 2 has had at least n_2 patients in it.

Plugging in our values of $b_1 = 10, b_2 = 2$ and $c_1 = 8, c_2 = 3$ gives us

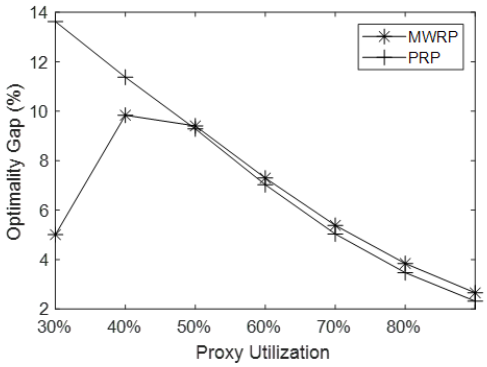
$$\frac{64}{45} \leq \frac{1}{\sqrt{n_2+1}-\sqrt{n_2}}. \tag{3.10}$$

We are now able to note that in equation (3.10), for all values of $n_2 \geq 1$, the minimum number of patients that need to be present in ward 1 (n_1) for an incoming patient to be directed to it is a negative value if there is at least 1 patient in ward 2 ($n_2 \geq 1$). This implies that the MWRP always routes an incoming patient to ward 1 and only routes patients to ward 2 once the capacity of ward 1 is met.

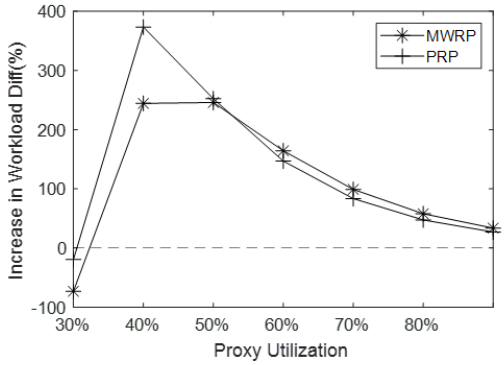
MWRP and ORP thus behave in an identical manner in this scenario where the ward with the higher number of nurses behaves in an IME fashion and the ward with the lower number of nurses behaves in a DME fashion. It must be noted here that this conclusion is heavily dependent on the chosen system parameter values b_1, b_2, c_1, c_2. A rigorous theoretical treatment is required before making any conclusive statements regarding the behavior of ORP or MWRP.

3.4.1.2 Scenario 2: Ward 1 DME; Ward 2 IME

Fig. 3.8 considers the scenario where ward 1 with more nurses has a DME form for patient time in service and nurse workload while ward 2 has the opposite. Unlike in Fig. 3.7, we see that MWRP does not appear to perform similarly to ORP. To understand this, we begin by recreating equation (3.10) with



(a) optimality gap measuring the percentage increase in long-run average patient time in system under heuristic when compared against ORP



(b) the percentage increase in long-run average difference under the heurstics when compared against ORP

Figure 3.8 Ward 1 concave; ward 2 convex. Both heuristics perform poorly in their attempt to reduce patient sojourn time and balance nurse workload compared to the ORP.

values and functional form considered in this scenario. Our corresponding relation is

$$n_2 \geq \frac{1}{2}\left(\frac{45}{64}\left(\sqrt{n_1+1}-\sqrt{n_1}\right)-1\right). \tag{3.11}$$

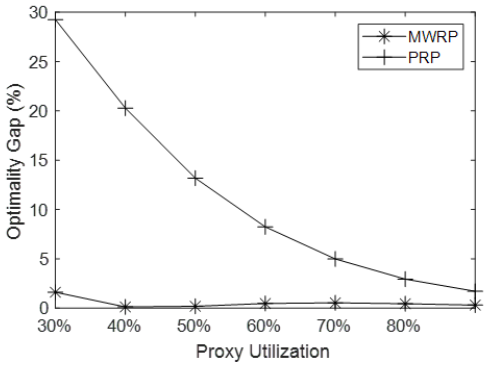
Following a similar exercise from equation (3.10), we are able to note that MWRP dictates filling up one ward (in this case ward 2) first before assigning patients to ward 1. In other words, MWRP dictates filling up the IME ward before assigning patients to the DME ward. However, ORP disagrees with this approach as ward 1, despite being concave, has more nurses than ward 2. As a result assigning patients to ward 1 even before filling up ward 2 is acceptable as the greater number of nurses are able to handle the sharp increase in average patient time in system arising due to the concavity of the ward. As a result, ORP is able to find values of β_1, β_2 that perform better than the MWRP.

We note here that the values of β_1, β_2 for ORP under this scenario did not display any discernible patterns like in scenario 1.

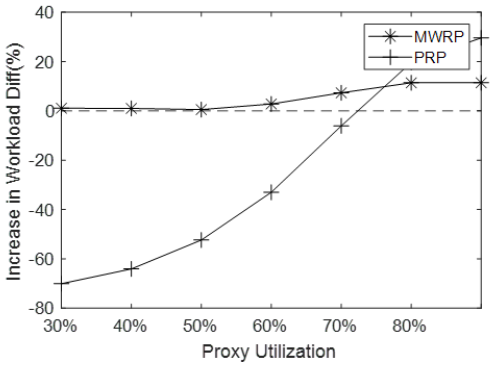
3.4.1.3 Scenario 3: Both Wards Concave

Fig. 3.9 considers the scenario where both wards 1 and 2 have a concave form for patient time in service and nurse workload. Furthermore, ward 1 has more nurses and patients spend longer in the system under max occupancy. Our first observation is that MWRP performs in a manner similar to ORP both in terms of optimality gap and percentage increase in workload difference at lower utilization levels. We are able

to confirm this observation by noting that the values of β_1, β_2 under ORP are close to -50, -50. Recall that -50 was used as a surrogate for $-\infty$. Our second observation is that though PRP performs worse than



(a) optimality gap measuring the percentage increase in long-run average patient time in system under heuristic when compared against ORP



(b) the percentage increase in long-run average difference under the heurstics when compared against ORP

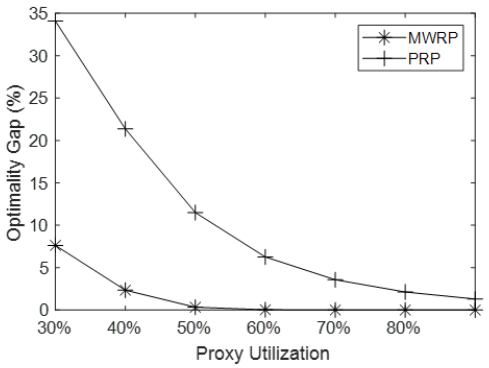
Figure 3.9 Both wards concave. MWRP performance is similar to ORP in reducing patient sojourn time at lower utilization levels but worse in balancing workload at higher utilization levels. PRP performs poorly in reducing patient sojourn time but balances workload much better than ORP at lower utilization levels.

MWRP and ORP at lower utilization levels in terms of optimality gap, it offers a substantial improvement over both MWRP and ORP in terms of reducing the difference in workload between wards. The reason for this can be understood by looking at how patient time in service changes in Fig. 3.6d. We can observe that even though both wards are concave, the extent of concavity is higher for ward 1 as the marginal increase in patient time in service (y-axis) is greater in the case of ward 1 at lower utilization levels. As a result, both MWRP and ORP tend to assign patients to ward 2 at lower utilization levels more often. While this improves a patient's time in system, the nurses in ward 2 end up with higher levels of workload. PRP on the other hand assigns patients in proportion to the number of nurses in both wards thus leading to a balance in workload between nurses of the two wards. Unfortunately, such as assignment by PRP comes at the cost of average patient sojourn time.

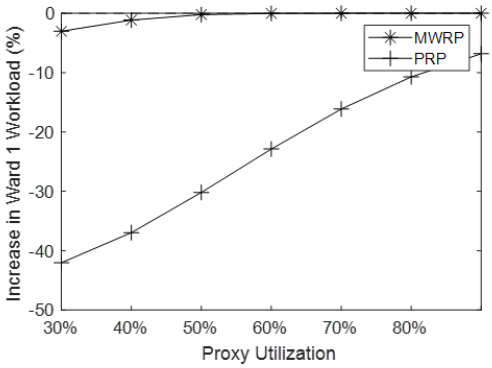
3.4.1.4 Scenario 4: Both Wards Convex

Fig. 3.10 considers the scenario where both wards 1 and 2 have a convex form for patient time in service and nurse workload. Our observations and conclusions about the performance of MWRP and PRP when

compared against ORP are very similar to those in scenario 3. ORP performance is similar to MWRP and better than PRP at all utilization levels in terms of minimizing a patient's time in system. However, PRP performs better than ORP and MWRP in terms of balancing workload at all utilization levels.



(a) Optimality gap of percentage increase in long-run average patient time in system under heuristic when compared against ORP



(b) percentage increase in workload of ward 1 under heuristic when compared against ORP

Figure 3.10 Both Wards Convex. The performance of MWRP is similar to ORP at higher utilization levels in terms of patient sojourn time and workload balance. PRP, while performing poorly in terms of sojourn time, performs much better in balancing workload than ORP at lower utilization levels.

3.4.2 Experiment 2: Comparison of ORP and MWRP Under Workload Constraints

Our second experiment studied the performance of ORP in the presence of workload constraints. The optimization model in Eq. (3.3) provides a decision maker with the flexibility to choose an appropriate constraint for nurse workload. In this experiment, we chose to balance workload between the nurses of the two wards by constraining our optimal policy to have a long-run average difference between nurse workload in the two wards to be under a certain threshold target γ^*. To determine the value of γ^*, we first obtained the long run difference in workload experienced by nurses in both wards under MWRP and set the value of γ^* to this vale multiplied by a threshold factor σ. In other words, we solved the optimization model with a constraint requiring the long-run average difference in nurse workload between wards to be under a proportion (σ) of the long-run average difference in nuse workload under MWRP. We then performed a series of experiments varying σ between 30% and 90% in increments of 10% while also varying the proxy system utilization between the same ranges with the same increments. We note here

that ORP was only compared to MWRP as the performance of the PRP was noted to be consistently worse than both MWRP and ORP during the first experiment in Section 3.4.1 as far as patient sojourn time is concerned.

Proxy Utilization

σ	30%	40%	50%	60%	70%	80%	90%
30%	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**
40%	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**
50%	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**
60%	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**
70%	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**
80%	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**
90%	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**

(a) Ward 1 convex; ward 2 concave.

Proxy Utilization

σ	30%	40%	50%	60%	70%	80%	90%
30%	**Inf**	-9%	-9%	**Inf**	**Inf**	**Inf**	**Inf**
40%	**Inf**	-9%	-9%	-7%	**Inf**	**Inf**	**Inf**
50%	**Inf**	-9%	-9%	-7%	**Inf**	**Inf**	**Inf**
60%	**Inf**	-9%	-9%	-7%	-5%	**Inf**	**Inf**
70%	-4%	-9%	-9%	-7%	-5%	-4%	**Inf**
80%	-4%	-9%	-9%	-7%	-5%	-4%	-3%
90%	-4%	-9%	-9%	-7%	-5%	-4%	-3%

(b) Ward 1 concave; ward 2 convex.

Proxy Utilization

σ	30%	40%	50%	60%	70%	80%	90%
30%	26%	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**
40%	22%	18%	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**
50%	22%	15%	12%	**Inf**	**Inf**	**Inf**	**Inf**
60%	13%	14%	10%	5%	**Inf**	**Inf**	**Inf**
70%	13%	9%	8%	3%	2%	**Inf**	**Inf**
80%	13%	5%	6%	2%	0%	1%	**Inf**
90%	4%	3%	3%	1%	0%	0%	0%

(c) Both wards concave.

Proxy Utilization

σ	30%	40%	50%	60%	70%	80%	90%
30%	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**
40%	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**
50%	9%	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**
60%	6%	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**
70%	5%	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**
80%	3%	4%	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**
90%	2%	2%	**Inf**	**Inf**	**Inf**	**Inf**	**Inf**

(d) Both wards convex.

Figure 3.11 Optimality gap for each scenario on varying σ and proxy system utilization $\hat{\rho}$. Note that values in each cell represent the increase in average patient sojourn time under ORP compared against MWRP.

Fig. 3.11 shows the results of our experiment that studies the performance of ORP with a workload constraint. Each of the four subfigures (a), (b), (c), and (d) provide an idea of how much worse the patient time in system gets on implementing ORP with an added constraint controlling the long-run average difference between nurse workload in two wards. Consider the matrix in Fig. 3.11a. We see that the model is infeasiable no matter what the proxy utilization is or the value of σ is. This is expected as we noted in Fig. 3.7a that when ward 1 is convex with more nurses and ward 2 is concave, MWRP is nearly optimal. As a result, any constraint that attempts to reduce the difference in workload between nurses beyond the value achieved by MWRP becomes infeasible. Now consider the matrix in Fig. 3.11b. We see that all feasible values are negative indicating that the constrained optimization problem finds a value for patient time in system lower than the corresponding value achievable by implementing MWRP.

Comparing against Fig. 3.8a provides us with a reason for the better performance of the constrained optimization problem as we see that MWRP performs worse than the ORP when ward 1 is concave and ward 2 is convex. We also noted in Fig. 3.8a that the performance of MWRP is worst around proxy utilization values of 40%-50%. This observation is reflected in Fig. 3.11b as around these values of proxy utilization, we see the least amount of in-feasibility. In other words, ORP is able to significantly reduce the average difference in nurse workload compared to MWRP while still achieving lower patient time in system. Finally we see in Fig. 3.11c and Fig. 3.11d that attempting to better balance nurse workload compared to MWRP always results in an increase in patient time in system. This indicates that if both wards are concave or convex, a decision maker must value the trade-off between balancing nurse workload and reducing a patient's total time in system.

3.4.2.1 Discussion About Experimental Analyses

Experiment 1 in Section 3.4.1 compares the performance of ORP against MWRP and PRP without workload constraints. We note here that the scenarios considered are by no means exhaustive and careful analyses must be done before making generalizeable conclusions. There may exist a variety of combinations for functional forms, function coefficients, and number of nurses that lead to unique situations and optimal solutions. The experiments we conducted appear to indicate that MWRP performs similar to ORP in most situations. The one scenario where MWRP performed poorly considered a concave functional form for patient time in service and nurse workload in the ward with a greater number of nurses. This is informative as it indicates that a decision maker could choose to increase the number of nurses assigned to a ward that appears to have a concave behavior in terms of patient time in service and nurse workload.

Experiment 2 in Section 3.4.2 informs the extent to which a decision maker can attempt to enforce constraints on nurse workload. In the four scenarios we considered, two increased a patient's time in system on attempting to balance nurse workload, one led to a completely infeasible model, and one case led to improved patient times in system on balancing nurse workload. This experiment reiterates the importance, from a decision makers perspective, about the importance of recognizing how nurses within a ward function on varying workload levels. However, it must be stressed again that careful data-driven analyses must be performed before arriving at conclusions.

3.5 Some Theoretical Considerations

3.5.1 Implications for MWRP in General Service Systems

Section 3.4 shows that implementing the MWRP often leads to near-optimal patient time in system. Our definition of MWRP is from a practical standpoint that focuses on sending incoming patients to the ward with the lowest marginal increase in workload on assignment. In this section, we briefly show that though our use of MWRP appears novel to service systems with workload considerations, its underlying principles have been commonly used in past service science literature even without accounting for workload.

Consider the probability $P_i(n_1, n_2, .., n_K, \bar{\beta})$ from equation (3.2) of assigning an incoming patient to ward i given as

$$P_i(n_1, n_2, .., n_K, \bar{\beta}) = \frac{\delta_i^{\beta_i}}{\sum_j \delta_j^{\beta_j}}$$

Recall that $\delta_i = \gamma_i(n_i + 1) - \gamma_i(n_i)$ corresponds to the marginal increase in workload experienced by the nurses in ward i on addition of a patient to the ward currently having n_i patients. Consider our general workload function $\gamma_i(n_i) = b_i \times f(n_i/c_i)$. The probability of assigning an incoming patient to ward i now equals

$$P_i(n_1, n_2, .., n_K, \bar{\beta}) = \frac{\Big(b_i \times f((n_i+1)/c_i) - b_i \times f(n_i/c_i)\Big)^{\beta_i}}{\sum_j \Big(b_j \times f((n_j+1)/c_j) - b_j \times f(n_j/c_j)\Big)^{\beta_j}} \tag{3.12}$$

Recall that MWRP is equivalent to setting $\beta_i \to -\infty \; \forall i$. This implies that equation (3.12) is equivalent to

$$P_i(n_1, n_2, .., n_K, \bar{\beta}) = \begin{cases} 1, \; i = \underset{i}{\operatorname{argmin}} \Big(b_i \times f((n_i+1)/c_i) - b_i \times f(n_i/c_i)\Big) \\ 0, \text{ otherwise.} \end{cases} \tag{3.13}$$

The policy outlined in equation (3.13) corresponds to a threshold policy. Each time a patient arrives, MWRP dictates that the patient be assigned to the ward i that has the smallest value of $b_i \times f((n_i + 1)/c_i) - b_i \times f(n_i/c_i)$, which can be expressed as a function of n_i. Finally, from equation (3.6), we note that n_i can be expressed as a function of $\mu(n_i)$. We thus note that MWRP dictates that the patient be assigned to ward i that has the smallest value of a function of $\mu(n_i)$. The MWRP thus bears resemblance to other such threshold policies widely studied in literature, notable among them being the $c\mu$ policy. We

note here that the existence of this threshold policy for MWRP exists under the specific system setting of the workload defined as a second-order polynomial function in n_i.

3.5.2 Analysis of ORP From the Perspective of Service Rate

The implications for patient routing throughout this chapter were considered from the lens of marginal increase in workload. In other words, the probability $P_i(n_1, n_2, .., n_K, \bar{\beta})$ from equation (3.2) of assigning an incoming patient to ward i has been defined in terms of the marginal increase in workload of a ward on patient assignment $\delta_i^{\beta_i}$ as

$$P_i(n_1, n_2, .., n_K, \bar{\beta}) = \frac{\delta_i^{\beta}{}_i}{\sum_j \delta_j^{\beta}{}_j}$$

Recall also that per our definition, $\delta_i^{\beta_i} = b_i f((n_i+1)/c_i) - b_i f((n_i)/c_i)$.

However, our definitions of workload and service rate mean that the two are linked and one can be expressed in the form of the other. Consider the mean service time $m_i(n_i)$ for a patient in ward i when there are n_i patients in the ward. From equation (3.5), $m_i(n_i)$ can be expressed as the inverse of the rate $\mu(n_i)$ as,

$$m(n_i) = \mu_i(n_i)^{-1} = a_i^{min} + (a_i^{max} - a_i^{min}) \times q\left(\frac{n_i/c_i - \varepsilon^{min}}{\varepsilon^{max} - \varepsilon^{min}}\right). \tag{3.14}$$

For ease of notation, we assume the value of ε^{min}, which is the smallest possible value of patient-nurse ratio, to equal 0. Next, combining equations (3.14) and (3.15), assuming the same functional form form f and q, and expressing m_i in terms of δ_i gives us the following relation,

$$\delta_i = b_i q(\varepsilon^{max}) \frac{m_i(n_i+1) - m_i(n_i)}{a_i^{max} - a_i^{min}}. \tag{3.15}$$

Expressing $\frac{b_i q(\varepsilon^{max})}{a_i^{max} - a_i^{min}}$ as T_i now gives us a simplified relationship between the marginal increase in workload on patient assignment δ_i and the average patient service time m_i

$$\delta_i = T_i(m_i(n_i+1) - m_i(n_i)). \tag{3.16}$$

We can re-write the above expression as

$$\delta_i = T_i \frac{m_i(n_i+1) - m_i(n_i)}{n_i + 1 - n_i} \equiv T_i \Delta m_i(n_i). \tag{3.17}$$

The last equivalence holds as 1 is the smallest unit of measure for $m_i(n_i)$ as a result of which $\frac{m_i(n_i+1)-m_i(n_i)}{n_i+1-n_i}$ is equivalent to the slope of the m_i curve at n_i. Substituting equation (3.17) into the probability function for patient assignment and re-writing $m_i(n_i)$ as μ_i^{-1} gives us

$$P(n_1, n_2, .., n_K, \bar{\beta}) = \frac{(T_i \Delta \mu_i^{-1})^{\beta_i}}{\sum_j (T_j \Delta \mu_j^{-1})^{\beta_j}}. \tag{3.18}$$

Equation (3.18) now takes a form similar to parametrized rate-based policies [Dor16], also known as r-routing policies. Under an r-routing policy, the next job in queue is assigned to idle server i with probability

$$P_i(\mu, r) = \frac{\mu_i^r}{\sum_j \mu_j^r} \tag{3.19}$$

where μ_i is the service rate of the ith server. For special values of the parameter r, we recover well-known policies. For example, setting $r = 0$ results in Random; as $r \to \infty$, it approaches Fastest Server First (FSF); and as $r \to -\infty$, it approaches Slowest Server First.

A notable difference between the r-routing policy in equation (3.19) and our routing policy in equation (3.18) is that while the r-routing policy focuses on making assignment based on the value of service rate, our policy focuses on making assignments based on the how the value of the service rate changes on assignment. This observation poses an interesting question for future research: how would a policy based on both absolute value and derivative of service rates perform? Addressing this question is, however, outside the scope of this dissertation.

3.5.3 Notes on Extending Model Beyond 2 Wards

The numerical analysis in Section 3.4 considered an ED with 2 wards. It is not uncommon however for emergency departments to have 3 or more wards and hence it is important for us to be able to analyze and solve our model with 3 or more wards. Such an extension to the model is accompanied by computational difficulties. Consider the optimization model in Eq. (3.3). A problem setting with n wards with a maximum of m patients in each ward and a maximum of w patients waiting leads to the constraint set 5.1 having $(w+1)(m+1)^n$ constraints. The inclusion of each additional ward in the model, thus, scales up the problem size by a factor of $m+1$. Furthermore, these equations are non-linear owing to the exponential form for the patient routing probability. We handled these numerical complications by

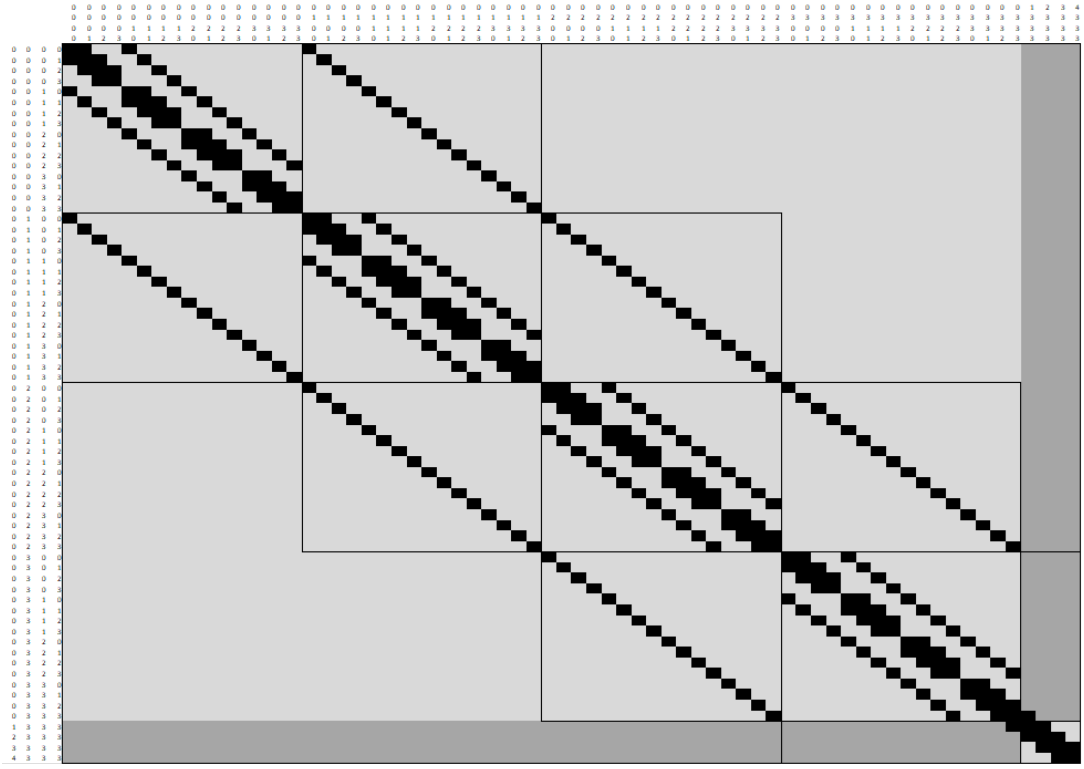


Figure 3.12 Structure of the rate transition matrix $Q(\bar{\beta})$ under a 3-ward setting with a maximum of 2 patients in each ward and a maximum of 3 waiting spots. A cell within the matrix $Q(\bar{\beta})$ is marked as black if it is non-empty and is marked white otherwise.

leveraging the special problem structure (that of a LDQBD process) as shown in Fig. 3.4. We see in Fig. 3.12 that extending the problem beyond 2 wards preserves the block diagonal structure as we defined the levels of our LDQBD via the number of patients present in the first ward. Per this definition, the index of levels can only go up or down by one, since the number of patients in a ward can only go up or down by one in a markovian process. An important point for consideration is the fact that we have not discussed what the structure of CTMC transitions within a level should be. In other words, we have not made any assumptions regarding structure of the generator matrices $A_0(i,\bar{\beta}), A_1(i,\bar{\beta}), A_2(i,\bar{\beta})$ outlined in Section 3.3.2. The structure of these matrices plays a significant role in determining the computational complexity of our proposed solution. For instance, defining the levels of the LDQBD process to be the number of patients waiting would lead to Q matrix being split up into exactly 4 generator matrices (shown via the varying shades of gray in Fig. 3.12). This would then lead to exactly one matrix inversion ($A_1(0,\bar{\beta})$, which is the top left gray matrix) being required and the computational difficulty would be determined based on the structure of this matrix. Furthermore, special matrix structures can still be

leveraged to perform this inversion. The example shown in Fig. 3.12 shows a block tri-diagnoal structure for $A_1(0,\bar{\beta})$ meaning that special algorithms [Meu92] can be used to obtain the matrix inverse.

We summarize this section with the following statements. Extending the model beyond 2 wards is relatively straightforward from a modeling standpoint as we maintain the LDQBD structure. However, we may run into computational difficulty depending on how the generator matrices of the LDQBD process is set up. A detailed investigation into the relationship between the structure of the generator matrices and the computational difficulty is outside the scope of this paper.

3.6 Conclusions

In this paper, we model patient flow through an ED where a patient's service time depends on the workload experienced by the nurses within the ward. We represent this workload as a function of nurse-patient ratio within the ward and develop a policy to route incoming patients into one of several wards based on the workload being experienced by nurses in each of the wards. The structural form of the optimal policy takes into account the marginal increase in workload on adding a new patient to each ward. The optimal policy is compared against two heuristics that (1) assign the incoming patient to the ward with the least marginal increase in workload on patient assignment (2) assign the incoming patient to wards in proportion to the number of nurses in the ward. Our results show that the optimal routing policy performs as well or better than a myopic minimum workload routing policy, depending on the nature of how workload and and average patient service time varies on increasing ward occupancy levels. Furthermore, we see that the optimal routing policy consistently performs better than a policy that routes patients to wards in proportion to the number of nurses in the wards.

A key takeaway from this paper is the importance of accounting for workload in developing strategies to route incoming patients to an ED. This paper assumes that a patient's time in the ED depends on the nurse-patient ratio of the unit. As a result, we see that the patient routing policy that attempts to balance the long-run workload between wards on assigning a new patient to a ward performs the best in terms of minimizing patient time in the system and fairness concerning the workload experienced by nurses in the wards.

We note that the model we developed in this paper may extend to other general service systems that we discussed as a part of our review of literature in Section 3.2. Some examples of such systems

include call centers and telecommunications networks. We contend that any service system that involves customer/agent arrivals into one of several different server pools can benefit from our work as the workload experienced by the servers plays an essential role in efficient provision of service. On the flip side, the optimal solution developed in this chapter is extremely granular and provides an explicit rule to route patients for every combination of number of patients in each ward at any point in time. Such a policy can be advantageous in certain settings. Routing policies in emergency departments are often heavily influenced by the lead/triage nurse who tends to have a holistic understanding of the mental, physical, and organizational state of the emergency department and the staffed nurses. It is important, thus, to provide the lead/triage nurses with the autonomy to route patients per their discretion due to sudden systemic changes (such as nurses falling ill, or a large sudden influx of patients) that are not easily captured by OR-based models. Granular models such as the ones developed in this chapter allow for this autonomy while ensuring that lead/triage nurse can return to using the optimal routing policy prescribed by this chapter once the sudden systemic issue has been resolved.

Finally, while the model developed in this paper may be generalized to multiple patient severity types and multiple wards, the methodology we use to find the optimal routing policy becomes computationally difficult as the problem size increases. This difficulty becomes especially exacerbated when one considers the important role that severity type plays in determining a patient's service rate. Such a consideration goes beyond hospitals and is true in several general service systems as well. Communication networks are good example where the time to process a network packet depends on the size of the packet. The 'service rate' of the packet, thus, depends on the size of the network packet which is a proxy for the 'importance' or 'severity' of the network packet. An avenue for future research lies in developing a computationally tractable method. A second avenue for future research lies in considering workload functions that depend on more factors than just the nurse-patient ratio. Workload experienced by nurses in a ward depends on a variety of factors aside from nurse-patient ratio such as patient arrival rates, patient severity, and time of day. Working these factors into the model could present a more realistic picture of the flow of patients within an ED.

CHAPTER

4

MANAGING EMERGENCY DEPARTMENT PATIENT FLOW AND NURSE STAFFING USING FLUID APPROXIMATIONS FOR MULTI-CLASS POOLED SERVICE QUEUES

4.1 Introduction

Efficient patient flow through an emergency department (ED) is a critical factor that contributes to a hospital's performance, which in-turn influences overall patient health outcomes. In this work, we model a multi-class many-server pooled queuing system where patients of different acuity levels receive care

from one of several nurse pools, each comprising an ED ward. We assume that a patient's time in service is a function of the ratio of nurses to patients in their unit. Our model reduces patient Length-of-Stay (LOS) and controls nurse workload by optimizing routing and nurse allocation decisions between the wards. First, we address the complexity of the queuing model control equations due to patients of multiple acuity levels present in the same ED unit with service rates that depend on patient acuity and unit workload. To do this we approximate the queuing system via a deterministic fluid model to describe the control equations via their first order behaviors and formulate an optimization model using these control equations.

A fluid queue is a mathematical approximation of a queuing model that treats the flow of customers similar to the flow of fluids. The earliest mentions of fluid queues referred to the concept as *Dam Theory*, due to the models being used to describe fluid levels in a reservoir [Mor56]. A fluid queue can essentially be thought of as a large reservoir that has fluid pouring into it via a network of pipes connected to it. These pipes are proxies for agent arrival into a reservoir which is a proxy for the queue. Fluid approximations to queues have been used by operations researchers to study the first-order behavior of complex queuing systems under heavy loads [Whi06b]. Fluid approximations to queues have also been studied in contexts not involving service systems to model the spread of wildfires [Gri17] and in ruin theory [Bad09]. A fluid queuing model typically arises by scaling up the arrival rate and number of servers leading to unique deterministic, continuous, linear fluid limits. These limits are then substituted back into the queuing model equations to obtain the corresponding fluid model equations. The primary advantage to studying a queuing model in its fluid limits is due to the fluid queuing model's ability to characterize all the equations in terms of their first-order behavior, thus removing the complexity associated with the stochasticity involved in the queuing network.

In this chapter, the results obtained from the fluid optimization problem are input to a simulation developed with AnyLogic software to obtain performance measures of interest for the non-approximated system performance measures of interest, such as patient LOS and unit workload. Simulation as a tool has seen widespread use in healthcare systems engineering and process improvement [Hal13] and has been helpful in allowing decision makers to carry out analyses such as (1) resource (human, equipment, and space) allocation, (2) determining key performance indicators that affect the healthcare system, (3) estimating and improving the robustness of a system under sudden and unexpected situations [Liu16]. Perhaps the most valuable asset that simulation provides decision makers is the ability to perform

"what-if" scenarios to test various interventions and understand the efficacy of implementing them in place of the status quo. Simulation thus allows us to reconstruct complex service systems, such as a hospital ED, and to study and improve the its related systems and processes. In this chapter, we simulate patient flow and nurse assignment in the emergency department of Southeastern Regional Medical Center (SRMC) located in Lumberton, North Carolina. We use retrospective data obtained from SRMC to model patient arrivals, patient assignments to wards, nurse assignment to wards, and patient length of stay. Once validated against existing data, we run various 'what-if' scenarios by re-routing patients and re-assigning nurses based on the information obtained from solving the fluid optimization model. We provide a detailed analytic treatment for each of the 'what-if' scenarios and discuss the merits and demerits of implementing them. Our results (1) highlight the importance of accounting for nurse workload and service behavior in developing routing/staffing policies and (2) show that small changes to patient routing policies could lead to reduced patient LOS and better-balanced nurse workloads.

An outline for the remainder of this chapter is as follows. Section 4.2 gives an overview of literature related to staffing and routing within multi-class multi-server problems and their applications in healthcare and general service systems. Furthermore, Section 4.2 provides a brief history of fluid queuing and simulation models and their use in literature. Section 4.3 then describes the emergency department setting that we abstract to form a queuing model followed by an approximation to form a fluid model. Furthermore, we describe the development of the simulation model and the various input functions and parameters used to build it using data from SRMC hospital. Finally, in Section 4.4, we conduct several case studies by optimizing patient routing and nurse staffing under different constraints and discuss the optimal solution and its implications under each case.

4.2 Review of Literature

We begin this section by reviewing past work related to the use of operations research tools to solve patient routing models under a multi-class framework in general service systems. For literature related to patient routing in healthcare we refer the reader to Section 3.2. Next, we delve into the history of fluid approximations to queuing models and study the various settings, healthcare or otherwise, where such approximations have been used. Finally, we provide a short overview of the use of simulation in addressing complex problems related to healthcare operations. We restrict our attention to simulation

models focusing on the emergency department and highlight those models that have optimal patient routing as one of their goals. We note here that although nurse workload is an important feature of our work, we refer the reader to the literature review in Section 2.3 and 3.3 in Chapters 2 and 3 respectively for details.

4.2.1 Routing Under Multi-Class Arrival Settings

Multi-class arrival models may be broadly sub-categorized depending on whether the servers are homogeneous or non-homogeneous and are often referred to as the 'V' or 'W' models respectively [Gar00]. The former considers multiple classes of customers arriving to a single pool of servers with similar service rate characteristics while the latter considers multiple server pools with distinct service rate characteristics. Call centers have seen some of the earliest and most extensive applications for the 'V' models with several surveys including reviews of methods to deal with staffing and scheduling [Koo06].

The primary differences between various routing rules in literature are due to the following system characteristics.

- The specific system assumptions under which the rule is being implemented such as server and/or customer type dependent service rates [Ata04].
- The performance metric that the rule is designed to optimize such as holding cost, or abandonment rate for queued customers [Lar14].
- The regime under which the system is being analyzed such as heavy traffic or the Halfin-Whitt regime [Har04b].

Koole & Pot [Koo06] discuss that one of the most straightforward ways of analyzing multi-class models in call center applications is using Markov Decision Processes [Put14]. It is theoretically possible to take all relevant factors under consideration as long as the state of the system defined for the purposes of the Markov process is memoryless. However, the size of the state-space can grow very large very quickly which leads to the need to devise simpler and more implementable policies. Hierarchical routing [Kle77] and priority routing [Wal04] were two commonly used policies with a focus on being simple both in definition and implementation. Hierarchical routing, also known as overflow routing, operated by assigning the incoming customer to the first available service agent, after traversing through all agents

according to a certain ranking of agents. Thus, the agent with the highest priority is considered first and is assigned to the customer if available, otherwise the agent with the second highest priority is considered. The objective of developing such a policy lies in determining the optimal priority ranking depending on the performance metric that one wishes to optimize. Hierarchical routing, while being simple to describe and requiring little storage data, suffers from the drawback of requiring that the priority order of agents be fixed prior to beginning the routing process. Priority routing is an extension of hierarchical routing where each incoming customer type may possess different priority rankings for the server agents. While priority routing policies have the advantage of satisfying both customers and agents, and having high flexibility by means of the control parameters (that determine the priority list for each customer group), Sisselman & Whitt [Sis07] discuss that the policy is incapable of identifying an optimal routing algorithm that takes into account server job satisfaction. Furthermore, neither hierarchical routing nor priority routing allows for a framework involving customers receiving service from an entire pool of agents.

Skill-Based routing (SBR) has recently gained popularity in the literature related to call centers [Wal05]. SBR assigns the incoming customer to the most suitable service agent as opposed to the next available service agent. What makes a service agent suitable depends upon characteristics of the problem being considered. Koole & Mandelbaum [Koo02] provide an excellent overview of the challenges surrounding the need for skill-based routing. They discuss that most call centers are multi-type multi-skill operations involving incoming customers of multiple types requiring service from call center agents with multiple skill sets. According to them, a common way of implementing skill-based routing is by specifying two selection rules: agent selection - how an arriving call selects an idle agent; and call selection - how an idle agent selects a waiting call. Wallace & Whitt [Wal05] state that most strategies to handle multi-type multi-skill operations are not optimal without the use of numerically infeasible algorithms like dynamic programming and discuss the need for fundamental principles and operations research techniques that make it possible to better design and manage SBR call centers. The authors go on to develop an algorithm to both route and staff in a SBR call center. Whitt [Whi06a] used fluid approximations to queues to determine optimal routing and staffing in a contact-center with SBR. The author defines contact-center as any system where a customer is required to come in contact with an agent to receive service. A call center is a type of contact center. Our work in this chapter borrows from many of the principles outlined by Whitt [Whi06a]. Two key differences between our works are (1) the pooling of service leading to state-dependent service rates and (2) the consideration for server (in our

case, nurse) workload. We will discuss the similarities and differences further in Section 4.3.

Finally, we note that the question of staffing is often closely related to routing as optimal staffing depends on the choice of routing rule used, and vice-versa. Most research related to the combined staffing and routing problem has focused on various forms of approximations. While patient flow within an ED has been widely studied by OR practitioners, [Hal06; Sag15; Zel11; McH11], Harrison & Zeevi [Har05] mention that staffing and routing are often treated in a separate but hierarchical manner due to the computational complexity involved. In other words, either staffing or routing is optimized first before fixing the corresponding optimal policy and optimizing the other.

The model developed in this chapter bears similarities to work done by many authors in the past. However, our work stands out due to the presence of all of the following salient features within the same framework.

- Patients receive service from an entire pool of nurses within a ward. This feature is similar to both Chapters 2 and 3 of this dissertation.
- Patient service time is dependent on both their own severity type, on the ratio of all patients to nurses of the ward they are in, and on the type of ward they are in.
- Optimal patient routing and nurse staffing are done in conjunction with consideration for nurse workload via constraints in the optimization model. Furthermore, we allow for a generalized workload constraint depending on the specific needs of the decision maker.
- Routing and staffing decisions are made in a combined fashion as opposed to the hierarchical fashion followed by many researchers, as outlined by Harrison & Zeevi [Har05].

4.2.2 Fluid Approximations to Queuing Models

The concepts involved in developing the equations of fluid queues are analogous to equations of the *leaky bucket algorithm*, which Turner [Tur86] is credited with first describing [Tan03]. While the leaky bucket algorithm deals with a deterministic source (fluid arrival rate), fluid approximations to queues are an application of the leaky bucket algorithm to a stochastic source. While early iterations of fluid queues took the form of 'dam theory', Anick et al. [Ani82] was one of the earliest authors to apply the concept of fluid approximations to an application area unrelated to actual fluid flows. Specifically, they modeled

the transmission of network packets in a telecommunications network following which the term 'fluid queue' was coined [Lat18].

The basic idea behind fluid queues is this. Consider a regular $M/M/s$ queuing system. Let us assume that the s servers belong to a 'server pool' and an incoming customer/agent is serviced by the first available server from the pool. An incoming customer that sees all servers busy has to wait in a 'buffer'. Let us assume that this buffer has infinite capacity. Now, suppose that the rate at which customers arrive is scaled up by a certain factor η. Concurrently, suppose that the size of the server pool is also scaled by the same factor. As the scaling factor $\eta \to \infty$, the rate at which customers arrive is so large that the inflow behaves like water flowing out of a faucet at a constant rate. Under this scaling, the pool of servers behaves like a reservoir with $\eta \to \infty$ droplets of fluids (customers) all being serviced at a constant rate. Here lies the primary advantage of analyzing a stochastic queueing process in its fluid limits; no matter the stochasticity associated with the arrival and service distribution in the original queueing system, under the fluid model these rates behave in a deterministic fashion. The fluid queue has been shown to be asymptotically correct in the scaled regime for the Markovian $M/M/s+M$ model [Man95; Whi04] and a discrete-time analog of the general $G_t(n)/GI/s+GI$ model [Whi06b]. However, Whitt [Whi06a] argues that the discrete-time setting may be used as an approximation for continuous-time cases as time increments of the discrete-time case can be arbitrarily short. This characteristic of fluid queues allows us to model complex arrival and service rate functions and to obtain tractable analytic solutions.

Past researchers have used fluid models to solve OR problems in healthcare. Yom-Tov & Mandelbaum [YT14] analyzed a queueing model with reentrant customers motivated by healthcare systems with patient recidivism. They validated their model against an ED and applied time-varying fluid approximations to propose a time-varying square-root staffing policy. Yousefi et al. [You19] proposed a scheduling procedure for outpatients requesting appointments at healthcare facilities. They defined the scheduling process as an MDP that tries to decrease the wait times for higher priority outpatients and used fluid approximiation to estimate the optimal solution to the MDP. In his thesis, Anderson [And14] considered the problem of scheduling medical residents in hospitals and used fluid approximations to queuing models to capture and analyse the insights obtained from departments operating at a critical capacity regime. Dotoli et al. [Dot09] proposed a model to describe the structure and dynamics of a critical ED and describes the complete workflow and management of patients starting from their arrival to the ED until either their or admission to a suitable department. They used a fluid approximation to define an

optimization problem to determine optimal resource allocations strategies.

In this chapter we first create a queuing model for patient flow through an ED. Next, we address the complexity of the queuing model control equations due to patients of multiple acuity levels present in the same ED unit with service rates that depend on patient acuity and unit workload. To do this we approximate the queuing system via a deterministic fluid model to describe the control equations via their first order behaviors. We borrow from the work of Whitt [Whi06a] to create the fluid model with the following notable differences.

- Patients in our model receive service from the entire pool of servers with a service time that is dependent on patient severity type and the ratio of nurses to patients of the unit that the patient is in. The model developed by Whitt [Whi06a] assumed that each customer receives service from exactly one service agent.

- Our model establishes constraints on the workload experienced by nurses within a ward. We accomplish this by establishing a measure for average number of patients of each severity type in every ward to estimate the fluid approximated value for nurse-patient ratio. The model developed by Whitt [Whi06a] did not consider server workload.

4.2.3 The Use of Simulation in Optimizing Emergency Department Flows

Simulation modeling has seen widespread use in the field of health systems engineering [Bre10]. A significant reason for this widespread use is the advancement in computational power and data storage capacity over the last two decades. The various paradigms of simulation modeling allows one to understand and study vastly different aspects of a system. For instance, while discrete event simulation allows us to characterize and analyse the flow of processes over time, agent-based modeling allows us to study all combinations of interactions between agents in a system and the system itself. Some examples of healthcare related operational problems where simulation modeling has had an impact include resource allocation in hospitals [Wen99], patient flow [Hal06], ambulance routing [McL13], hospital bed utilization [Dum85], and epidemiology [Dav19].

The operational complexity of an Emergency Department often necessitates the use of simulation in performing a quantitative analysis of patient flow, wait times, and factors that influence patient outcomes. One of the earliest models simulating an ED was developed by Saunders et al. [Sau89]. They developed a

computer simulation model of an ED. The model incorporated multiple levels of patient priority, assigned each patient to an individual nurse and physician, and incorporated tests, procedures, and consultations. Patient throughput time in the model varied directly with lab service times and inversely with the number of physicians or nurses. The simulation model was descriptive in nature and was aimed at showing animations of patients, and staff members moving throughout an ED. More recently, a review by Connelly & Bair [Con04a] showed that Discrete-Event-Simulation (DES) has been the most common paradigm of simulation employed in modeling an ED. There is a vast amount of literature pertaining to the use of DES in modeling and improving emergency department operations. We refer the reader to review papers by Günal & Pidd [Gün10], Thorwarth & Arisha [Tho09], and Wiler et al. [Wil11] for a more comprehensive overview.

In this chapter we develop a model to simulate patient flow through SRMC's emergency department. We use real hospital data to obtain patient arrival rates, patient service times, patient routing, and nurse staffing. The purpose of the simulation is twofold. The first is to validate input data from SRMC against patient length of stay and staffing estimates. The second purpose is to use the simulation in a manner similar to what was done by Dotoli et al. [Dot09] where the authors used a simulation model to show that the policy resulting from the optimal decision parameters obtained from the fluid model is more efficient than the status quo. Finally, we use the simulation to perform experiments describing how patient LOS and nurse workload varies when implementing different scenarios of optimal decision variables.

4.3 Model Development

The analytic treatment within this chapter is divided into four major pieces, each outlined over the upcoming subsections. Section 4.3.1 defines a mathematical abstraction of patient flow through an ED by developing a queuing model. Section 4.3.2 introduces a fluid approximation to the queuing model before using the fluid model to define an optimization model. Section 4.3.3 describes the estimation of various input parameters such as arrival rates and patient service time from data available from SRMC's ED. Section 4.3.4 details the development of the simulation model using AnyLogic software.

4.3.1 Queueing Model

In this section we define a multi-class queuing model to represent the arrival and service process within a hospital emergency department. Let us consider patients of I different severity types and J wards, with ward j, $\forall\ 1 \leq j \leq J$ containing s^j nurses. Each ward j can house a maximum of M^j patients which may be greater or less than the number of nurses s^j. This is because we assume that the patients within a ward receive pooled service from all the nurses in the ward. Unlike a traditional queuing model, we don't assume that a single patient is served by a single nurse.

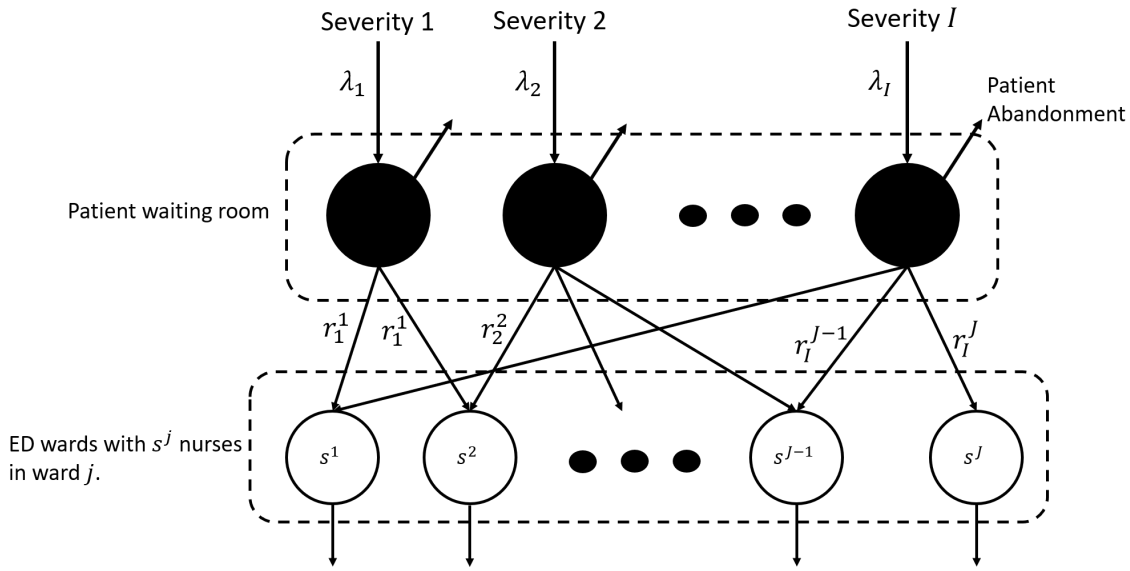


Figure 4.1 An overview of the patient flow process being considered in this Chapter. Patients of different severity levels arrive to an ED and are assigned to a ward if space is available and depending on the given routing policy. If there is no space, patients wait until they are able to join a ward. Patients waiting for too long may abandon the system before entering service in a ward.

A pictorial representation of the patient flow process is provided in Fig. 4.1. Patients of severity type i arrive to the system with an exponentially distributed inter-arrival time with arrival rate λ_i. These patients may be assigned to ward j according to a routing proportion r_i^j with $\sum_{j=1}^{J} r_i^j = 1$ for $1 \leq i \leq I$. In other words, a proportion r_i^j of patient of severity i are served by nurses in ward j. These proportions may be thought of as probabilities and are treated as decision variables in our model. We assume that each patient severity type is associated with a queue where they wait if they are unable to enter service immediately upon arrival. We assume an infinite buffer for this queue. After arrival and before joining

service, patients may abandon (leave after joining the queue but before starting service). We assume that successive times to abandon for patients of severity type i are i.i.d random variables with CDF F_i.

We assume that the time spent in service by a patient is a function of the number of nurses in the ward and the number of patients of all different severity types in the ward. In other words, the random variable S_i^j corresponding to the amount of time a patient of severity type i spends in service in ward j depends on number of nurses s^j in ward j and the number of patients $\mathbf{n}^j = (n_1^j, .., n_I^j)$ of all severity types $i \in \{1, .., I\}$ in the ward. We characterise this dependency by assuming the mean value of this random variable $E[S_i^j]$ to be the function m_i^j as

$$E[S_i^j] \equiv \mathrm{m}_i^j(s^j, \mathbf{n}^j) \rightarrow R_{>0}.$$

Here, we do not specify the form of the function m_i^j, this is discussed further in section 4.3.3. An important feature of our model is the ability to optimize staffing and routing while ensuring that the workload experienced by nurses in wards is kept under predefined thresholds. We assume that the workload experienced by nurses in ward j depends both on the number of nurses in the ward s^j and the number of patients of each different severity type $\mathbf{n}^j = (n_1^j, n_2^j, ..n_I^j)$ according to the function γ_j as

$$\gamma_j : (s^j, \mathbf{n}^j) \rightarrow R_{\geq 0}$$

We now specify a general workload constraint

$$\Psi(\gamma^j(s^j, \mathbf{n}^j), \forall, j) \in \psi$$

in a manner similar to the formulation in Eq. (3.3) in Chapter 3. We state that a desired function Ψ of the workload of all the wards be within a desired range of set ψ. While performing experiments, we consider two types of workload constraints. The first attempts to keep the workload of each ward under a pre-defined threshold while the second attempts to keep the absolute difference in threshold across all pairs of wards under a pre-defined balance threshold.

Our model has two decisions to be made: staffing and routing. Staffing is the choice of numbers s^j for $1 \leq j \leq J$ that specifies how many nurses must be assigned to each ward while routing is the choice of numbers r_i^j for $1 \leq i \leq I$ and $1 \leq j \leq J$ that specifies the ward that incoming patients of a specific

severity type must be assigned to. Characterizing the queuing model described so far via closed-form expressions is difficult. In the next section, we will define an approximating fluid model that allows us to characterize the system via a set of equations.

4.3.2 Fluid Model

We approximate our queuing model by scaling up the arrival rates of patients, ward capacity, nurse staffing, while fixing abandonment distribution, and service time distributions. In order to scale, we follow the procedure outlined by Whitt [Whi06a] who introduced a family of models indexed by a scaling parameter η, and then let $\eta \to \infty$. The arrival rates, maximum patient capacity in a ward, and number of servers are then set to be functions of η as

$$\frac{\lambda_i(\eta)}{\eta} \to \lambda_i \ , \ \frac{M^j(\eta)}{\eta} \to M^j \ \text{ and } \ \frac{s^j(\eta)}{\eta} \to s^j \ \text{ as } \ \eta \to \infty.$$

Thus, $\lambda_i(\eta) \approx \eta\lambda_i$ is the arrival rate of patients into the queuing model η but λ_i is the arrival rate of class-i fluid after scaling. Similar interpretations hold for $M^j(\eta)$ and $s^j(\eta)$.

Our fluid model is characterized by the parameter seven-tuple $(\lambda, \mathbf{x}, \mathbf{F}, \mathbf{r}, \mathbf{S}, \mathbf{s})$ where $\lambda \equiv (\lambda_1, \lambda_2, .., \lambda_I)$ is an I-tuple of numbers corresponding to arrivals, $\mathbf{F} = (F_1, ..., F_I)$ is an I-tuple of CDFs corresponding to abandonment, $\mathbf{S} \equiv (S_i^j : 1 \leq j \leq J, 1 \leq i \leq I)$ is an $I \times J$ matrix of service time CDFs, $\mathbf{x} \equiv (x_i^j : 1 \leq j \leq J, 1 \leq i \leq I)$ is an $I \times J$ matrix of numbers corresponding to number of patients of each severity type in a ward, $\mathbf{r} \equiv (r_i^j : 1 \leq j \leq J, 1 \leq i \leq I)$ is an $I \times J$ matrix of numbers corresponding to patient routing proportions, and $\mathbf{s} \equiv (s_1, .., s^j)$ is a J-tuple of numbers corresponding to ward staffing.

To describe how the fluid model evolves in time, we define w_i as a deterministic time a fluid of class i waits before entering service. This measure is relevant as the proportion of customers who do not abandon while waiting for service equals $F_i^c(w_i)$ (the CCDF of the abandonment distribution after class i fluid has waited for time w_i).

We are now in a position to define the system control equations for our fluid model. This is where our model begins to see significant departures from the framework outlined by Whitt [Whi06a]. We begin by recognising that fluids of two different classes within a ward do not interact. This is because the fluids of two different classes are able to share the same pool of nurses at the same time. This is unlike in a traditional queuing system where if one of the servers was occupied due to serving a particular class of

fluid, that server is unavailable to other fluid classes. As a result, the effect of the number of servers s^j is seen only in the service time function m_i^j.

Before we define the system control equations, we note that since service time in our original queuing model is defined by the number of patients, we require a scaling of the number of patients and need to represent it by a certain amount of fluid. Thus, if n_i^j is the number of patients of type i in ward j in the queuing model, we define x_i^j as its scaled counterpart in the fluid model as

$$\frac{n_i^j(\eta)}{\eta} \to x_i^j \text{ as } \eta \to \infty.$$

The service time function $\mathrm{m}_i^j(s^j, \mathbf{x}^j)$ can thus be defined as a fixed deterministic quantity since both s^j and $\mathbf{x}^j$ are fixed deterministic numbers.

We can now express the system control equations for each ward in terms of the system control equations for each fluid type within the ward. In other words, we can express the system control equations via the expression 'rate-in = rate-out' for each class i fluid in ward j. Now, the arrival rate of class i fluid entering service at ward j (which is also the 'rate in') equals $\lambda_i r_i^j F_i^c(w_i)$. The first term ($\lambda_i$) is the overall arrival rate of fluid i. The second term (r_i^j) is the proportion of fluid i that is routed to ward j while the last term ($F_i^c(w_i)$) is the proportion of class i fluid that does not abandon after having waited for w_i units of time. The mean service time for class i fluid entering ward j equals $\mathrm{m}_i^j(s^j, \mathbf{x}^j)$. The rate out thus equals the inverse of the mean service time ($\mathrm{m}_i^j(s^j, \mathbf{x}^j)^{-1}$) multiplied by the service capacity (x_i^j) giving us the following control equation::

$$\lambda_i \times r_i^j \times F_i^c(w_i) = x_i^j \times \mathrm{m}_i^j(s^j, \mathbf{x}^j)^{-1}, \ \forall i, \forall j.$$

In addition we have the following sets of constraints to prevent fluid loss during routing

$$\sum_i r_i^j = 1, \ \forall \ 1 \leq j \leq J.$$

Finally, we have a set of constraints to ensure that the total amount of fluid (of all classes) is capped in each ward j according to maximum capacity M^j as

$$\sum_i x_i^j \leq M_j \ \forall \ 1 \leq j \leq J.$$

4.3.2.1 Optimization Model

Before defining the objective of our optimization model, we first define the associated cost and reward coefficients. The reward functions and coefficients we outline are a direct adaptation from [Whi06a] with the exception that we ignore cost associated with balking. We assume a positive reward rate $v(i,j,t)$ earned per unit of fluid per unit time for serving class-i fluid in ward j after customers have waited for time t. This reward decreases in t. Next, we assume a cost $c^a(i,t)$ incurred per unit of time for fluid of class i that abandons after waiting for time t. Finally, we assume a holding cost of $c^h(i,y)$ incurred per unit time for having y units of class-i fluid waiting in queue. y in the expression $c^h(i,y)$ is the amount of class-i fluid waiting in queue in the fluid limit and is calculated using the expression $\lambda_i \int_0^{w_i} F_i^c(t)\mathrm{dt}$. We thus have the following expression for total reward that we wish to maximize.

$$R \equiv R(\mathbf{s},\mathbf{r},\mathbf{w}) = \sum_i \left(\lambda_i F_i^c(w_i) \sum_j r_{i,j} v(i,j,w_i) - \lambda_i \int_0^{w_i} c^a(i,t)\mathrm{d}F_i(t) - c^h\left(i, \lambda_i \int_0^{w_i} F_i^c(t)\mathrm{dt}\right)\right).$$

We note here that the above expression does not explicitly minimize a patient's overall LOS. However, by attempting to reduce wait times with abandonment penalties, the model incentivizes the system to establish smart routing and staffing policies that lead to faster patient service. This in turn ensures that patient wait time is reduced further downstream in the wait queues.

The complete optimization model may now be written as follows

$$
\begin{aligned}
\underset{\mathbf{w},\mathbf{r},\mathbf{s}}{\text{maximize}} \quad & \sum_i \left(\lambda_i F_i^c(w_i) \sum_j r_{i,j} v(i,j,w_i) - \lambda_i \int_0^{w_i} c^a(i,t)\mathrm{d}F_i(t) - c^h\left(i, \lambda_i \int_0^{w_i} F_i^c(t)\mathrm{d}t\right)\right) \\
\text{subject to} \quad & \lambda_i r_i^j F_i^c(w_i) \mathrm{m}_i^j(s^j, \mathbf{x}^j) = x_i^j, \;\; 1 \le j \le J, \le i \le I, && (1) \\
& \sum_j r_i^j = 1, \;\; 1 \le i \le I, && (2) \\
& \sum_i x_i^j \le M_j, \;\; 1 \le j \le J, && (3) \\
& \Psi(\gamma^j(s^j, \mathbf{x}^j), \forall, j) \in \psi, && (4) \\
& 0 \le \mathbf{r} \le 1, && (5) \\
& \sum_j s^j = \Theta, && (6) \\
& 0 \le \mathbf{w}, && (7)
\end{aligned}
\tag{4.1}
$$

Here, Θ is the maximum number of servers available for assignment. We have not placed any restriction on the nature of the functions $\gamma^j(s^j, \mathbf{x}^j)$ and $\mathrm{m}_i^j(s^j, \mathbf{x}^j)$. Presumably, $\gamma^j(s^j, \mathbf{x}^j)$ would increase with an increase in the amount of fluid ($\mathbf{x}^j$) in the ward and would decrease with the addition of servers (s^j). We do not place any restrictions on the functional form of the increase or decrease. Similarly, we do not place any restrictions on the form of $\mathrm{m}_i^j(s^j, \mathbf{x}^j)$.

We note here that additional constraints may be included depending on a decision makers requirement. An example of such constraints would be to specify a minimum number of nurses required in any given ward. Another example would be to specify that some patient severity types not be routed to specific wards.

4.3.2.2 System Control Analysis

Constraint set (1) in Eq. (4.1) can be used to establish necessary conditions and assumptions about system stability. However, the conditions are highly dependent on the form of the function $\mathrm{m}_i^j(s^j, \mathbf{x}^j)$. We will demonstrate in this section the procedure to establish conditions pertaining to system stability, minimum staffing, and required minimum patient wait time under the assumption of linearity for function $\mathrm{m}_i^j(s^j, \mathbf{x}^j)$. A similar procedure may be followed for higher order polynomial functional forms. We note

here however that a higher-order polynomial form is considered while inferring m_i^j function from real data in Section 4.3.3.

Consider constraint set (1) in the optimization model. Specifically, let us assume that the average service time function takes the following polynomial form

$$m_i^j(s^j, x^j) = a_i^j + b_i \times \frac{\sum_i x_i^j}{s^j},\ 1 \leq j \leq J. \tag{4.2}$$

Here, a_i^j is the baseline average amount of time (with no other patients in the ward) it takes a patient of severity i in ward j to complete service. b_i is the factor by which each additional patient being cared by a nurse on average in the ward increases the average time it takes for a patient of severity i to complete service and exit the ward. Constraint set (1) may now be written as

$$\lambda_i r_i^j F_i^c(w_i)\left(a_i^j + b_i \times \frac{\sum_i x_i^j}{s^j}\right) = x_i^j,\ 1 \leq j \leq J \tag{4.3}$$

Rewriting $\lambda_i r_i^j F_i^c(w_i)$ as $\hat{\lambda_i}^j$ to represent the effective arrival rate of patients of severity type i to ward j we have

$$\hat{\lambda_i}^j\left(a_i^j + b_i \times \frac{\sum_i x_i^j}{s^j}\right) = x_i^j,\ 1 \leq j \leq J \tag{4.4}$$

Summing up both sides of the equation across patient type i and rearranging terms gives us the following relationship

$$\frac{\sum_i a_i^j \hat{\lambda_i}^j}{1 - \frac{\sum_i b_i \hat{\lambda_i}^j}{s^j}} = \sum_i x_i^j,\ 1 \leq j \leq J. \tag{4.5}$$

We know that the $\sum_i x_i^j \leq M^j$ in order for ward j to not exceed the maximum capacity. This gives us the following inequality

$$\frac{\sum_i a_i^j \hat{\lambda_i}^j}{1 - \frac{\sum_i b_i \hat{\lambda_i}^j}{s^j}} \leq M^j,\ 1 \leq j \leq J. \tag{4.6}$$

Rearranging in terms of of the inequality gives us the following system stability conditions,

$$\frac{\sum_i a_i^j \hat{\lambda_i}^j}{M^j - M^j \frac{\sum_i b_i \hat{\lambda_i}^j}{s^j}} \leq 1,\ 1 \leq j \leq J. \tag{4.7}$$

Equation (4.7) provides us with a lower bound for the number of nurses required in any given ward.

Rearranging in terms of s^j gives us the following expression for minimum number of nurses required to staff each ward;

$$s^j \geq \frac{\sum_i b_i \hat{\lambda_i}^j}{1 - \frac{1}{M^j}\sum_i a_i^j \hat{\lambda_i}^j},\ 1 \leq j \leq J. \tag{4.8}$$

We obtain the wait time thresholds for patients of severity type i by rearranging the terms in equation (4.8) to obtain an expression in terms of $\hat{\lambda_i}^j$. We then simplying the equation successively as follows,

$$s^j - \frac{s^j}{M^j}\left(a_i^j \hat{\lambda_i}^j + \sum_{k\neq i} a_k^j \hat{\lambda_k}^j\right) \geq b_i^j \hat{\lambda_i}^j + \sum_{k\neq i} b_k^j \hat{\lambda_k}^j,\ 1 \leq j \leq J,$$

$$b_i^j \hat{\lambda_i}^j + \frac{s^j a_i^j \hat{\lambda_i}^j}{M^j} \leq s^j - \sum_{k\neq i}\left(\frac{s^j a_k^j \hat{\lambda_k}^j}{M^j} - b_k^j \hat{\lambda_k}^j\right),\ 1 \leq j \leq J,$$

$$\hat{\lambda_i}^j \leq \frac{M^j s^j + \sum_{k\neq i} \hat{\lambda_k}^j \left(M^j b_k^{je} - s^j a_k^{je}\right)}{M^j b_i^{je} + s^j a_i^{je}},\ 1 \leq j \leq J,$$

$$F_i^c(w_i) \leq \frac{M^j s^j + \sum_{k\neq i} \hat{\lambda_k}^j \left(M^j b_k^{je} - s^j a_k^{je}\right)}{\lambda_i r_i^j \left(M^j b_i^{je} + s^j a_i^{je}\right)},\ 1 \leq j \leq J,$$

$$F_i(w_i) \geq 1 - \frac{M^j s^j + \sum_{k\neq i} \hat{\lambda_k}^j \left(M^j b_k^{je} - s^j a_k^{je}\right)}{\lambda_i r_i^j \left(M^j b_i^{je} + s^j a_i^{je}\right)},\ 1 \leq j \leq J,$$

$$w_i \geq \min_{1\leq j\leq J}\left\{F_i^{-1}\left(1 - \frac{M^j s^j + \sum_{k\neq i} \hat{\lambda_k}^j \left(M^j b_k^{je} - s^j a_k^{je}\right)}{\lambda_i r_i^j \left(M^j b_i^{je} + s^j a_i^{je}\right)}\right)\right\}. \tag{4.9}$$

Equation (4.9) thus provides us with the minimum amount of time a patient of severity type i will have to wait depending on other system parameters including the routing proportions and wait times for other patient severity types

4.3.3 Input Parameters

To fully describe our model, we require the following five sets of operational parameters estimates that characterize patient flow: (1) arrival rate by patient severity type (λ_i), (2) current nurse staffing levels for each ward during the status quo (s^j), (3) maximum ward capacity (M^j), (4) CDF for patient abandonment for each patient severity type (F_i), (5) routing proportion of patient severity type to each ward (r_i^j), and (6) rate function for the time spent by a patient in service (μ_i^j). We note here that any mention of status quo

henceforth in this Chapter refers to the current set of operational parameters being used in the hospital. Our goal is to modify and optimize the parameters corresponding to nurse staffing (s^j) and patient routing (r_i^j).

We begin by describing the data available to us. We obtained input for our model by inferring from over 88,000 unique visits, each with over 150 variables including timestamps, visit attributes and patient outcomes. Patient level visit data was available from November 2017 to April 2019 while physician and nurse schedules along with daily bed assignments were available for 2018 and 2019. There exist patients of 5 different severity types (with 1 being the most severe and 5 being the least severe) and 3 different wards that the patients may receive service in. The wards are named critical care, minor care, and fast track. The critical care ward is typically occupied by patients of severity 1 and 2 while the fast track ward is usually visited by less severe patients. We refer the reader to work done by our colleague, Swan et al. [Swa19], for a more detailed description of the data and how certain parameters like nurse-patient ratios were estimated. In our work we borrow estimates for nurse staffing, maximum ward capacity, and nurse-patient ratio (which we use to estimate patient service time distribution, described later) from Swan et al. [Swa19].

We inferred patient arrival rate by calculating the inter-arrival times for each patient severity type and assuming that these times came from a homogeneous exponential distribution. Correspondingly, we fit the inter-arrival time data to an exponential distribution using the package 'fitdist' in R programming language. Fig. 4.2 gives the values for arrival rates, staffing levels, maximum capacity and routing proportions that we obtained from data and used in our numerical analyses.

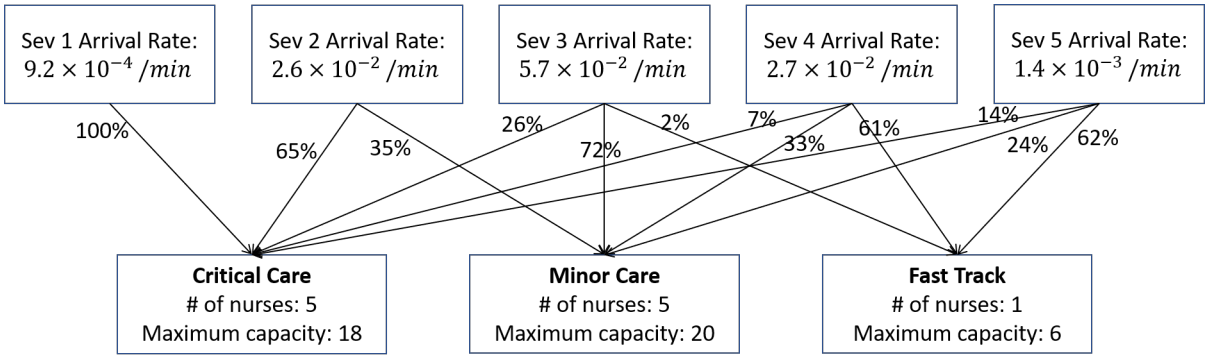
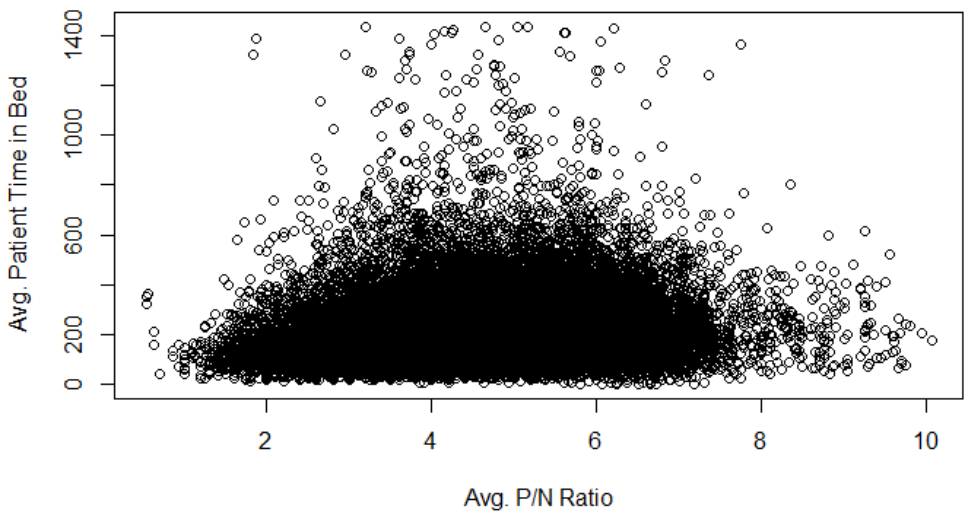


Figure 4.2 Pictorial representation for the status quo values of patient arrival rates, routing proportions, ward staffing, and ward capacity at SRMC's ED.

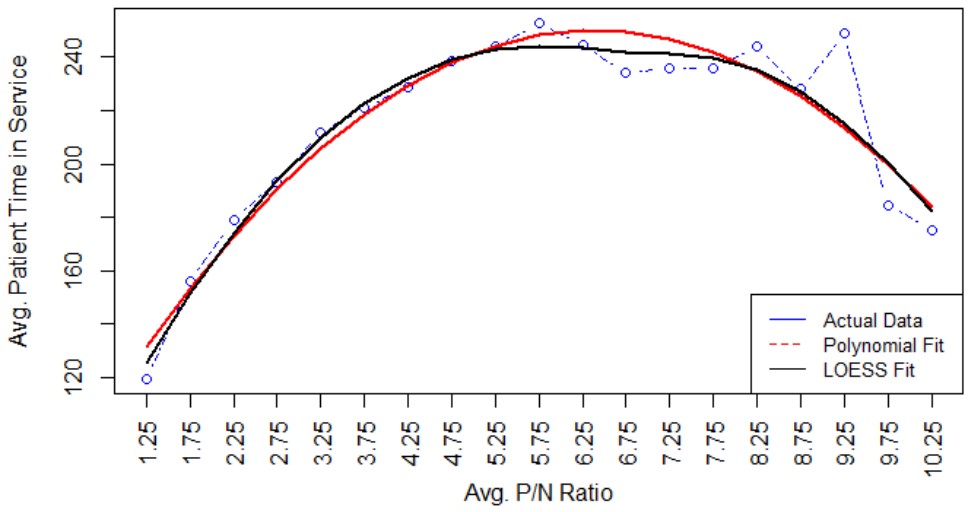
Estimating the CDF for patient abandonment from data is not trivial, due to the hospital being unable to keep records of when a patient abandons. Though the data had a small percentage of patient departures from the system coded as LWBS (left without being seen), this number refers to those patients who, after triage, had been assigned to a bed but departed before being seen by a nurse or physician. The lack of sufficient data meant that we assumed patients were unlikely to abandon in our model unless they waited for an extremely long period of time. We thus assumed in our model a weibull function for patient abandonment distribution with high values for scale and shape parameters (assumed to be the same and equal to be 18, 14, 13, 12, 10 for patient severity types 5, 4, 3, 2, and 1 respectively). Such a distribution ensured that the probability of patient abandonment remained nearly zero unless a patient waited for a long time. For instance, the probability of patients of severity type 5 abandoning becomes greater than 10% only after waiting 8 hours. This probability peaks to the maximum possible of around 35% at 10 hours before going back to be under 10% after 11.5 hours. We note here that patient wait times were never that high in the fluid model or simulation and therefore patients were never likely to abandon as a result of using this distribution.

The final and arguably most complex input parameter estimate is the function determining how long a patient stays in a ward. Our goal was to model a patient's time in service as a function of nurse-patient ratio. While there may exist a number of ways of estimating this function, we decided to use the time-averaged patient-nurse ratio (inverse of nurse-patient ratio) during a patient's stay in the ward for each patient's stay to fit a function relating the patient's total time in ward to the average patient-nurse ratio during the patient's stay. Details about how the the time-average for patient-nurse ratio is calculated is outlined in the work by Swan et al. [Swa19].

A scatter-plot of the average patient-nurse ratio against a patient's time in service is shown in Fig. 4.3b for patients of severity type 3 (moderate severity level) in the minor care ward. It must be noted here that we excluded data points with LOS values greater than 24 hours to remove outliers. We are able to make an immediate observation about the extreme spread of the scatter-plot rendering it difficult to fit a function. However, Fig. 4.3b shows a modified version of the scatter-plot. Essentially we divide up the values on the x-axis (average patient-nurse ratio) into buckets and calculate the average value of patient time in service (y-axis) within each of those buckets. We see that the curve now follows a distinct polynomial form. We are able to fit the curve in Fig. 4.3b to a second order polynomial function for the relationship between average patient time in service (y) and the average patient-nuse ratio



(a) Scatter-plot of average patient time in service against average patient-nuse ratio



(b) Average patient time in service fit against bucketed values of average patient-nurse ratio showing actual data, a LOESS fit, and a polynomial fit that we used in our experimental analyses.

Figure 4.3 Plots showing the steps in determining a functional form the relationship between average patient time in service and average patient-nurse (P/N) ratio for patients of severity type 3 in the minor care ward.

(x) - $y = 106.9 + 25.3x - 1.121x^2$. Fig. 4.3b also shows the curve-fit from LOESS (Locally Estimated Scatterplot Smoothing) regression which is a popular tool used to create a smooth line to better help

visualize the relationship between the variables [Cle91]. We see that the LOESS fit matches up quite well with the polynomial fit in this case. Finally, we undertake a similar procedure for all combinations of patient severity types and ward they're in. The resulting polynomial equations for each patient severity type within each ward is provided in Table 4.1. We note here that some combinations for patient type and ward are listed as N/A since there was not enough data to infer any sort of a functional form. These included patients of severity type 1 in minor care and fast track wards and patients of severity type 2 in the fast track ward. Accordingly, we restrict these routing combinations in our optimization model while performing experimental analyses. In other words, we enforce constraints that force the model to prevent patients of type 1 from going to minor care and fast track wards, and patients of type 2 from going to fast track wards.

Table 4.1 Mean patient service time categorized by patient severity type and ward. *pn* in the above expressions refers to patient-nurse ratio of the ward.

Severity Type	**Ward**	**Mean Service Time** (m_i^j)
1	Critical Care	$185.1 - 6.54pn + 0.86pn^2$
2	Critical Care	$140.37 + 22.52pn - 0.87pn^2$
3	Critical Care	$84.63 + 28.06pn - 1.05pn^2$
4	Critical Care	$113.07 + 3.78pn - 0.09pn^2$
5	Critical Care	$21.52 + 14.97pn - 0.29pn^2$
1	Minor Care	N/A
2	Minor Care	$224.05 + 39.05pn - 2.25pn^2$
3	Minor Care	$107.81 + 25.39pn - 1.12pn^2$
4	Minor Care	$43.75 + 19.70pn - 0.77pn^2$
5	Minor Care	$56.46 + 10.48pn - 0.05pn^2$
1	Fast Track	N/A
2	Fast Track	N/A
3	Fast Track	$149.97 + 1.5pn - 0.05pn^2$
4	Fast Track	$76.55 + 7.55pn - 0.28pn^2$
5	Fast Track	$172.33 + 25.46pn - 1.5pn^2$

It is interesting to note the form of Fig. 4.3b. We see that a patient's time in service first increases on increasing patient-nurse ratio as each nurse in the ward is required to care for more patients on average. However, at higher values of patient-nurse ratio, the patient time in service begins to decrease. We contend that this decrease is due to the nurses attempting to provide faster service due to the knowledge of the high workload that they are facing with so many patients to care for. However, the data does not

provide us with any information about patient recidivism or outcomes. It is likely that patients who are cared for under high patient-nurse ratio values return to the hospital or experience worse outcomes despite departing initially after a smaller time spent in the ward. Furthermore, we note here our assumption is similar to the one made in Chapter 3 that patient outcomes are not dependent on the time spent by them in the system. In other words, we do not account for long-term patient health outcomes or whether a patient after discharge left the ED to go home or was admitted to the hospital.

Finally, we note that a linear function was assumed for workload by separating the patient-nurse ratio terms by each patient severity type. Thus, the workload function took the form

$$\gamma_i(s^j, \mathbf{n}^j) = \sum_i u_i^j \frac{n_i^j}{s^j}.$$

While performing numerical experiments we assumed similar workload coefficients u_i^j across wards and only assumed difference across patient severity types. We did this to allow for easier representation of the workload measure when attempting to keep it under desired thresholds or to balance it across wards. However, we note that differences in patient health outcomes across wards are captured by variation in coefficient values for patient service time for patients of the same severity type across different wards, as seen in Table 4.1 The values used for u_i during the experiments in Section 4.4 are $10, 8, 7, 3, 2$ for $i = 1, 2, 3, 4, 5$ respectively, indicating that more severe patients lead to a higher level of workload for the nurses.

4.3.4 Simulation Model

We developed our simulation using AnyLogic software's personal learning edition. Each new agent is generated from one of 5 different source modules (one for each severity type) with an inter-arrival time distributed exponentially with a rate value that's shown in Fig. 4.2. If all the delay modules (representing wards) are at capacity, the patient enters a queue module (representing the wait room). On entry to the queue module, a random variable is drawn from the CDF for the patient's abandonment distribution. Once a patient has waited in the queue module for an amount of time equal to the drawn random variable, the patient is pushed out of the module and the counter for abandonment of the patient's severity type is incremented by one.

Patients are routed to wards according to predefined routing proportions obtained either from Fig. 4.2

or from solving the fluid optimization model. Once a patient enters a ward, the time that they will spend in the ward is determined by drawing a random variable from an exponential distribution with rate function that is dependent on the patient's severity type, the ward that the patient is in, and the patient-nurse ratio of the ward. Recall that this rate function is estimated using the scatter-plot of patient-nurse ratio versus patient time in service shown in Fig. 4.3. It must be noted here that each time a patient enters or leaves a ward, the simulation draws a new random variable for the patient's remaining time in service. This allows us to effectively capture the memory-less property of the state-dependent exponential distribution that we assumed for a patient's time in service.

Output statistics that we keep track of include average patient LOS and average ward workload. The average patient LOS includes the time spent by a patient waiting in queue and the time in service. The average ward workload is obtained per minute by summing up the workload of the ward calculated using the equation provided in Section 4.3.3 across the model's time horizon of one year and dividing by the total number of minutes in one year. We note here that the model's time horizon of one year only begins after completing a warm-up period (set as two weeks in our simulation).

4.3.5 Solution Procedure top Optimize Routing and Staffing

The full solution procedure used to obtain optimized values for patient LOS and ward staffing is outlined as follows.

- First, we solve the fluid optimization problem in equation (4.1). We note here that constraint (4) corresponding to the workload constraint may be modified to according to the experiment being considered. Examples of such modifications are provided while performing experimental studies in Section 4.4.2.
- The solution of the fluid optimization problem, specifically the routing proportions and staffing levels r_i^j and s^j, is then input to the simulation model in AnyLogic software.
- We run multiple replications of the simulation model and store the value of average patient LOS and ward workload for each replication.
- We finally record the result for optimal patient LOS and ward workloads as the average across all the replications.

We note here that in step 2 above, the solution for staffing levels s^j from the fluid optimization model is a continuous value. However, we require integer values for staffing in the simulation model. In order to generate the integer values, we consider all possible combinations for floor and ceiling values of the non-integral s^j such that the total number of nurses is the same. We ignore those combinations that lead to infeasibility. Out of all the feasible combinations, we calculate the objective value for each and select the combination with the best objective value.

Finally, we are also able to note that integer solutions for staffing levels may not be necessary as staffing levels can be represented as a fractional value of Full Time Equivalent (FTE) hours. We are able to observe (results not shown) while performing experiments that those policies that do not rely on heuristics to obtain integral staffing values lead to better weighted patient LOS and ward workload values. However, all of the experimental results we show use the aforementioned heuristic procedure that determines integer staffing values to allow for ease of interpretation of staffing results.

4.4 Numerical Analyses & Optimal Strategies

All of the experimental analyses were performed on a 8×1.6GHz Windows 10 machine with 8 GB of RAM. The fluid optimization was conducted on Matlab software using the 'interior-point' algorithm of the fmincon function. A description of the algorithm is provided by MATLAB [MAT]. We do not report run-times as all the optimization runs completed in under 10 seconds and all the simulation runs completed in under 20 seconds. In the next two subsections, we begin by first validating the fluid model before performing multiple experiments on a case study of the performance of SRMC's ED showing the implementation of our hybrid fluid optimization/simulation framework.

4.4.1 Fluid Model Validation

Before analysing the solutions of the fluid optimization model, it is important to validate that the control equations representing the fluid model match the real scaled system. In order to do this, we first formulate a system of equations corresponding to equation set (1) of Eq. (4.1) using the system parameter values in Fig. 4.2 and Table 4.1. Recall that the arrival rate, ward capacity, and nurse staffing in the fluid model need to be scaled up by a factor η. Accordingly, we parametrize the system of equations using the fluid scaling factor η. Next, we solved the system of non-linear equations using the 'fsolve' function (using

the trust-region dogleg algorithm Conn et al. [Con00]) in Matlab and obtained a set of solutions, one for each value of $\eta = 1, 2, .., 20$. For each solution set, we track the value of $x_i^{j-fluid}$ representing the amount of type-i fluid in ward j (or in other words, the number of patients of severity i in ward j in the scaled system).

Next we run the simulation model multiple times by scaling up the arrival rates, ward capacities, and ward staffing levels by values of $\eta = 1, 2, .., 20$ during each run. On completion of each run of the simulation, we recorded the average number of patients x_i^{j-sim} of each severity type $i \in I$ in every ward $j \in J$. For each value of η, we ran 30 replications of the simulation model and obtained an average across the 30 replications for each set of x_i^{j-sim} values. Finally, we computed the percentage deviation between the average fluid system state and the average simulation system state as

$$\frac{100\%}{|I| \times |J|} \sum_{i \in I,\, j \in J} \frac{|x_i^{j-fluid} - x_i^{j-sim}|}{x_i^{j-sim}}$$

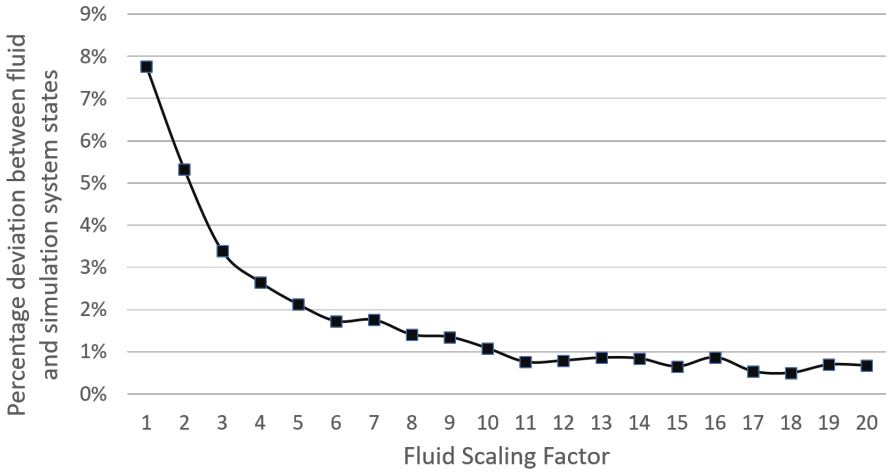


Figure 4.4 Percentage deviation in system state between fluid and simulation models on increasing the value of η.

Fig. 4.4 shows us that the percentage deviation between fluid and simulation systems states decreases to under 1% beyond η values of around 10. The implication here is that with a fluid scaling factor of around 10, the fluid control equations can be used as a good approximation for the actual simulation. We show in Section 4.4.2 how the optimal solution obtained from solving the fluid optimization model can be

approximated to provide us with an efficient solution for routing and staffing in the unscaled simulation representing the queuing model.

We note here that more work is required before determining the appropriate value of η for different system settings. Past researchers including [Hon15; Liu10; Sim95] have studied how well a fluid model mimics the real system on varying the value of η. Liu & Whitt [Liu10] showed the value of 20 to be a good lower bound for η in their work and demonstrated that stochastic deviations within their scaled queuing system is extremely small beyond this value of η. However, to the best of our knowledge, all studies involving fluid approximations consider a constant value for mean service rate with one-to-one service between customer and server. Recall that constraint set (1) in optimization model 4.1 establishes the system control equations for the fluid approximation. For a service system involving constant mean service rates and one-on-one interactions between server and customer, the equation transforms into

$$\lambda r^j F^c(w) \mu^j = s^j, \forall j$$

where μ^j represents the mean service rate and is a constant. For ease of notation, we consider a single customer type (fluid class) and drop the index i. As is clear from the above expression, scaling the arrival rate λ and number of servers s^j by the same value η cancels out on both sides of the equation. As a result, the value of η does not play a role in determining system state values in the fluid limit. The importance of η in such a situation comes from its ability to provide engineering confirmation to show that the family of scaled fluid systems matches a scaled real queuing system as $\eta \to \infty$.

This discussion tells us that choosing different values for η in a system such as ours involving non-linear mean service rates and pooled service can lead to different solutions for the fluid control system of equations as η on either side of the equation may not cancel out. This observation establishes the need for further study into how to choose an appropriate value for η. During the experimental analyses in this chapter, we consider a value of $\eta = 10$ as Fig. 4.4 suggests a deviation of under 1% between the scaled simulated system and the scaled fluid system states for $\eta > 10$. We note here that our objective in solving the fluid optimization problem is to obtain a possible routing and staffing policy which is then fed into a simulation to model actual patient flow and to obtain their corresponding length of stay. While choosing an appropriate value of η could provide us with policies that perform better, a relevant study of this choice of η is outside the scope of this dissertation.

4.4.2 Analyses of Optimal Solution Strategies

Having established that a fluid scaling value η of 10 leads to a reasonable match between the fluid and simulation models, we are now in a position to perform experiments on a case study concerning SRMC's Emergency Department. We consider the status quo parameter values for SRMC's ED as shown in Fig. 4.2 and Table 4.1. Next, we optimize the status quo routing and staffing values under different experimental settings that aim to control for workload by either balancing it between wards or by minimizing it across wards. Under each experimental scenario, we follow the solution procedure from Section 4.3.5 to obtain optimal values for patient LOS and ward workload. In addition to LOS for each patient severity type, we compute the weighted average for LOS across all severity types. We use arrival rates as the weights for each patient severity type.

Recall here the addition of a few constraints based on observations from data such as 1) patients of severity type 1 only being admitted to critical care and not minor care or fast track, and 2) patients of severity type 2 only being admitted to critical care and minor care and not to fast track. These constraints allow us to demonstrate special problem structure that may be necessary from an implementation standpoint, such as ensuring that patients of higher severity are not sent to a ward with nurses who aren't equipped to handle them.

We perform experimental studies to show the flexibility offered by the model to a decision maker in setting their desired operations goals for the system. The experiments are the following.

- **Optimization w/o workload constraints**: Here, we solve the fluid optimization model in Eq. (4.1) without workload constraint (4).

- **Optimization w/ workload threshold constraints**: Here, we solve the fluid optimization model in Eq. (4.1) with workload constraint (4) formulated to keep the long-run average workload of each ward under a pre-defined threshold (γ_j^*). The objective here is to minimize patient LOS while ensuring that ward workload does not exceed the specified thresholds. Specifically, the constraint set is formulated as

$$\gamma^j(s^j, \mathbf{x}^j) \leq \gamma_j^*, \; \forall \, j = 1, 2, 3.$$

  While performing the experiment, the value of γ_j^* is kept the same for all three wards and is equal to 15. We provide intuition behind this number by stating that under the workload function defined

by us in Section 4.3.3, the workload of a ward having one patient of each severity type being cared for by two nurses equals 15.

- **Optimization w/ workload balance constraints**: Here, we solve the fluid optimization model in Eq. (4.1) with workload constraint (4) formulated to keep the difference in workload between any two pairs of wards (j,k) under a pre-defined threshold ($\hat{\gamma_{jk}}$). The objective here is to balance the workload across wards while also reducing patient LOS. Specifically, the constraint set is formulated as

$$|\gamma^j(s^j,\mathbf{x}^j) - \gamma^j(s^j,\mathbf{x}^j)| \leq \hat{\gamma_{jk}},\ \forall,\ (j,k) \in \{(1,2),(2,3),(1,3)\}$$

  While performing the experiment, we kept the value of $\hat{\gamma_{jk}}$ the same for all three pairs of wards. We tested two different values of 5 and 2.5. To provide some intuition behind these numbers, let us consider two wards with one patient of each severity type in each ward. A difference in workload value of 5 between the two wards would mean that one ward has 3 nurses while the other has 2. Similarly, a difference in workload value of 2.5 would mean that one ward has 4 nurses while the other has 3. A difference in ward workload of 2.5 is more restrictive than 5 since the number of nurses is limited and in order to satisfy the workload balance constraint, the model would need to increase the number of nurses in all the wards.

4.4.2.1 Optimization w/o Workload Constraints

Solving the optimization model without any workload constraint gives us the opportunity to see the extent to which patient LOS can be reduced in the model. Table 4.2 shows a reduction in patient LOS for patients of all severity types by just modifying the routing proportions. The optimal policy suggests keeping the same staffing level as the status quo. We note from Table 4.2b that the the average patient LOS under the optimal parameters is reduced by 17% (35 minutes) on average with patients of severity type 5 seeing the biggest reduction by 44% and patients of severity type 1 the only ones seeing an increase in LOS by less than 1%. Furthermore, the critical care and minor care wards saw a reduction in workload while the fast track ward saw a large increase (230%) in workload. The optimal policy determined that in order to reduce the LOS for all severity types, it was necessary to route more patients of severity type 3 to the fast track ward (32% instead of 2%) which was the primary reason for the large increase in workload

Table 4.2 Optimal decision variables and performance measures on solving the optimization model without workload constraints.

(a) Optimal patient routing and nurse staffing decision variables compared against the status quo

Status Quo Routing and Staffing						
	Staffing	Severity 1 routing	Severity 2 routing	Severity 3 routing	Severity 4 routing	Severity 5 routing
Ward 1	5	100%	65%	26%	7%	14%
Ward 2	5	0%	35%	72%	33%	24%
Ward 3	1	0%	0%	2%	61%	62%
Optimized Routing and Staffing (Default Objective w/o workload constraints)						
	Staffing	Severity 1 routing	Severity 2 routing	Severity 3 routing	Severity 4 routing	Severity 5 routing
Ward 1	5	100%	57.2%	24.7%	3.8%	61.5%
Ward 2	5	0%	42.8%	43.3%	56.7%	19.8%
Ward 3	1	0%	0%	31.9%	39.6%	18.8%

(b) Optimal patient LOS and ward workload compared against the status quo.

Patient LOS (min)						
	Severity 1	Severity 2	Severity 3	Severity 4	Severity 5	Average
Status Quo	226	284	222	132	133	213.64
Optimized Unconstrained	228	250	193	82	75	178.32

Ward Workload			
	Critical Care	Minor Care	Fast Track
Status Quo	13.7	20.1	1.8
Optimized	12.1	10.9	29.1

for the nurse in the fast track ward. Finally, we note here that the value of the objective function improved by 0.8% on using the optimized routing and staffing parameters instead of the status quo parameters. This indicates that the existing policy being implemented at SRMC is close to optimal under our objective function.

4.4.2.2 Optimization w/ Workload Threshold Constraints

Table 4.2 showed us that a 10% (22 minutes) reduction in patient LOS is possible by modifying the routing proportions. However, it led to a significant imbalance in the resulting workload experienced by the ward nurses with nurses in the fast track ward seeing a major spike in workload.

Our next experiment attempts to control such an increase in workload by enforcing a constraint in Eq. (4.1) requiring that the long-run average workload in any ward be less than the value of 15. Table 4.3

Table 4.3 Optimal decision variables and performance measures on solving the optimization model with a constraint requiring the workload all wards to be under a value of 15.

(a) Optimal patient routing and nurse staffing decision variables compared against the status quo

Status Quo Routing and Staffing						
	Staffing	Severity 1 routing	Severity 2 routing	Severity 3 routing	Severity 4 routing	Severity 5 routing
Ward 1	5	100%	65%	26%	7%	14%
Ward 2	5	0%	35%	72%	33%	24%
Ward 3	1	0%	0%	2%	61%	62%
Optimized Routing and Staffing (Default Objective w/ workload threshold constraints)						
	Staffing	Severity 1 routing	Severity 2 routing	Severity 3 routing	Severity 4 routing	Severity 5 routing
Ward 1	3	100%	61.2%	11.9%	1%	10%
Ward 2	6	0%	38.8%	47.5%	39%	19.7%
Ward 3	2	0%	0%	41.6%	60%	70.3%

(b) Optimal patient LOS and ward workload compared against the status quo.

Patient LOS (min)						
	Severity 1	Severity 2	Severity 3	Severity 4	Severity 5	Average
Status Quo	226	284	222	132	133	213.64
Optimized Threshold	191	277	189	121	54	191.36

Ward Workload			
	Critical Care	Minor Care	Fast Track
Status Quo	13.7	20.1	1.8
Optimized	14.9	13.7	10.1

shows a slightly lesser reduction in patient LOS by enforcing workload threshold constraints for patients of all severity types by modifying both the routing proportions and the nurse staffing. The average reduction in patient LOS across all severity types is now 20% with patients of severity type 5 seeing the greatest reduction of 60% and patients of severity type 2 seeing the least reduction of 2.4%. We note here that even though the average reduction in patient LOS is lower than without enforcing any workload constraints, all the patient severity types see a reduction in LOS. Recall that under the experiment in Table 4.2, patients of type 1 did not see any significant change in LOS. Note that the optimal policy required that one nurse from the critical care ward be moved to the fast track ward in order to accomplish this.

4.4.2.3 Optimization w/ Workload Balance Constraints

Table 4.3 showed us that a reduction in patient LOS is possible while also controlling the workload by modifying both the routing proportions and nurse staffing. Our next objective was to attempt to balance the workload better among ward nurses instead of trying to control it to be under a threshold. To this end, we enforce a constraint in Eq. (4.1) requiring that the difference between average workload experienced by nurses in all pairs of wards be kept under a value of 2.5.

Table 4.4 Optimal decision variables and performance measures on solving the optimization model with a constraint requiring the absolute difference in workload between any two pairs of wards be under a value of 2.5.

(a) Optimal patient routing and nurse staffing decision variables compared against the status quo

Status Quo Routing and Staffing						
	Staffing	Severity 1 routing	Severity 2 routing	Severity 3 routing	Severity 4 routing	Severity 5 routing
Ward 1	5	100%	65%	26%	7%	14%
Ward 2	5	0%	35%	72%	33%	24%
Ward 3	1	0%	0%	2%	61%	62%
Optimized Routing and Staffing (Default Objective w/ workload balance constraints under 2.5)						
	Staffing	Severity 1 routing	Severity 2 routing	Severity 3 routing	Severity 4 routing	Severity 5 routing
Ward 1	4	100%	56.1%	24.4%	1.6%	83.5%
Ward 2	4	0%	43.9%	45.2%	58.4%	8.3%
Ward 3	3	0%	0%	30.4%	40%	8.2%

(b) Optimal patient LOS and ward workload compared against the status quo.

Patient LOS (min)						
	Severity 1	Severity 2	Severity 3	Severity 4	Severity 5	Average
Status Quo	226	284	222	132	133	213.64
Optimized Balanced-a	260	247	379	92	118	275.22

Ward Workload			
	Critical Care	Minor Care	Fast Track
Status Quo	13.7	20.1	1.8
Optimized	12.5	12.7	13.1

Table 4.4 shows us that while the optimal solution is able to achieve the objective of balancing the nurse workload across all three wards, the patient LOS sees an increase for some patient severity types, notably for severity type 3 (increase of 70%). Furthermore, severity type 1 sees an increase in LOS of

about 15%. The resulting average LOS across all patient severity types thus, is increased by 62 minutes (about 30%).

Table 4.4 suggests that the model is unable to establish an efficient routing and staffing policy when trying to keep the difference in workload between wards to under 2.5. However, relaxing the balance requirement and attempting to satisfy an average difference between ward workload to be under a value of 5 provides better results. Table 4.5 shows us that the average patient LOS sees a decrease of about 12 minutes (6%) in addition to satisfying the balance in workload constraint. The maximum difference in ward workload equals 4.7 and is seen between the minor care and fast track wards.

The key takeaway from this experiment is that attempting to balance ward workloads within the required operational parameters could lead to reduction in patient LOS. However, care must be taken in determining the balance threshold values. From a managerial perspective, this implies that a decision maker attempting to force too much of a balance could end up seeing negative effects from the point of view of patient LOS.

4.4.2.4 Summary of Results with an Assessment of Trade-offs

Section 4.4.3 shows the results of the experimental studies and show the flexibility offered by the model to a decision-maker in setting their desired operations goals for the system. We now summarize these results in this subsection by observing the trade-off between reducing patient LOS and balancing workload. We use aggregate performance measures for patient LOS and ward workload to study this trade-off. The aggregate measure we use for patient LOS is the weighted LOS across all patient severity types, the value of which we provide in the results table of each experimental study. For ward workload, we consider workload balance computed by calculating the average of the absolute value of difference in workload between all pairs of wards.

Fig. 4.5 summarizes the results of Section 4.4.3 by outlining the trade-offs involved in attempting to reduce patient LOS and increasing the workload balance between wards. We note that the policy that leads to the most substantial reduction in patient LOS (optimization with no workload constraint) only increases the workload balance levels by a small margin. We also note that the policy that leads to the highest increase in workload balance (optimization with tight workload balance constraint) is the only policy that leads to a substantial increase (of approximately one hour) in patient LOS. Fig. 4.5 thus appears to indicate that the policies that enforce a workload threshold constraint or a loose workload

Table 4.5 Optimal decision variables and performance measures on solving the optimization model with a constraint requiring the absolute difference in workload between any two pairs of wards be under a value of 5.

(a) Optimal patient routing and nurse staffing decision variables compared against the status quo

Status Quo Routing and Staffing						
	Staffing	Severity 1 routing	Severity 2 routing	Severity 3 routing	Severity 4 routing	Severity 5 routing
Ward 1	5	100%	65%	26%	7%	14%
Ward 2	5	0%	35%	72%	33%	24%
Ward 3	1	0%	0%	2%	61%	62%
Optimized Routing and Staffing (Default Objective w/ workload balance constraints under 5)						
	Staffing	Severity 1 routing	Severity 2 routing	Severity 3 routing	Severity 4 routing	Severity 5 routing
Ward 1	4	100%	56.1%	24.6%	2.6%	74%
Ward 2	4	0%	43.9%	44.1%	56.8%	17%
Ward 3	3	0%	0%	31.3%	40.6%	9%

(b) Optimal patient LOS and ward workload compared against the status quo.

Patient LOS (min)						
	Severity 1	Severity 2	Severity 3	Severity 4	Severity 5	Average
Status Quo	226	284	222	132	133	213.64
Optimized Balanced-b	242	241	241	85	82	201.52

Ward Workload			
	Critical Care	Minor Care	Fast Track
Status Quo	13.7	20.1	1.8
Optimized	11.8	12.5	17.2

balance constraint may be the most appropriate in achieving a trade-off between reducing patient LOS and balancing ward workload. However, we note here that Fig. 4.5 provides a decision-maker with information that helps them determine the objective most vital to them, allowing them to then to pick an appropriate joint routing and staffing policy.

4.4.3 Effect of Increasing Total Number of Available Staff

In all of the experimental scenarios that we considered, we fixed the total number of nurses available for allocation at a value of 11 in order to effectively compare our optimization results against the status quo. In this subsection, we consider the effects of increasing the total available staff on patient LOS. We chose the experimental scenario that attempted to keep the workload of each ward under a value of 15 as a representative base joint routing and staffing policy against which to compare increased staffing

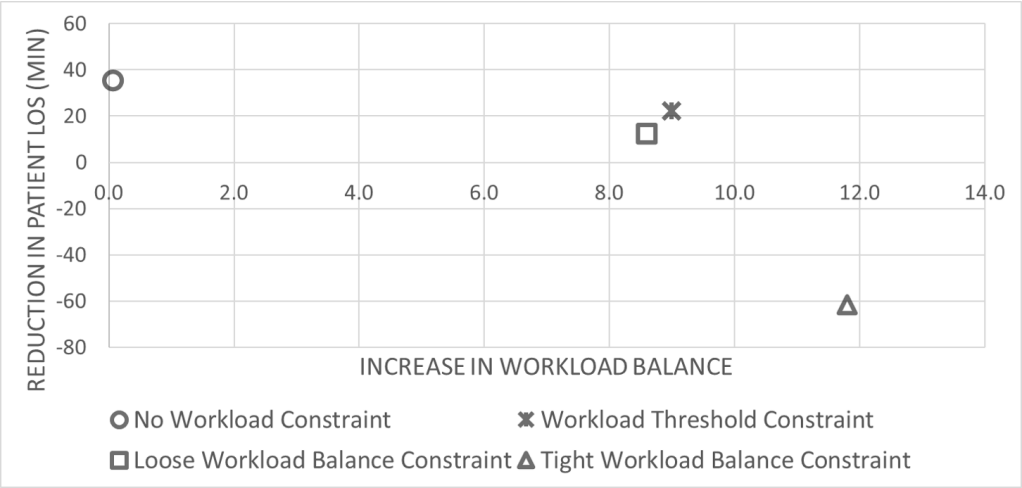


Figure 4.5 A comparison of the trade-off between patient LOS and ward workload balance. Here, loose and tight workload balance constraints refer to the experimental scenarios where we performed the optimization with a constraint attempting to keep the difference in workload between all pairs of wards to be under 5 and 2.5 respectively.

scenarios. Recall that we outlined the values for patient LOS and ward workload corresponding to this policy in Table 4.3b. First, we obtained optimal routing and staffing values from the fluid optimization model, assuming a higher number of total available nurses (specifically, we considered 12, 13, 14, and 15 nurses). We then obtained the simulated values for patient LOS under the optimal routing and staffing policy obtained under each scenario of increased staffing level. To compare against the status quo, we calculated the percentage reduction in weighted patient LOS under each scenario of increased staffing.

4.4.4 Discussions and Managerial Implications

During the course of this chapter we conduct multiple experimental studies to observe how the optimization framework developed by us performs under real-world conditions. The choice of which constraint to use is often left up to the discretion of the decision maker. However, we were able to show that despite the differences in constraint set leading to different routing patterns and staffing recommendations, the value of the objective remained fairly constant. This implies firstly that the current routing and staffing decisions being made by SRMC are already close to being optimal, from the point of view of the objective function which aims to minimize waitime while reducing patient abandonment as a result of long waittimes. Making small changes to routing and staffing could lead to a reduction in patient LOS of nearly 17% while also ensuring that the workload experienced by nurses across wards is kept under a

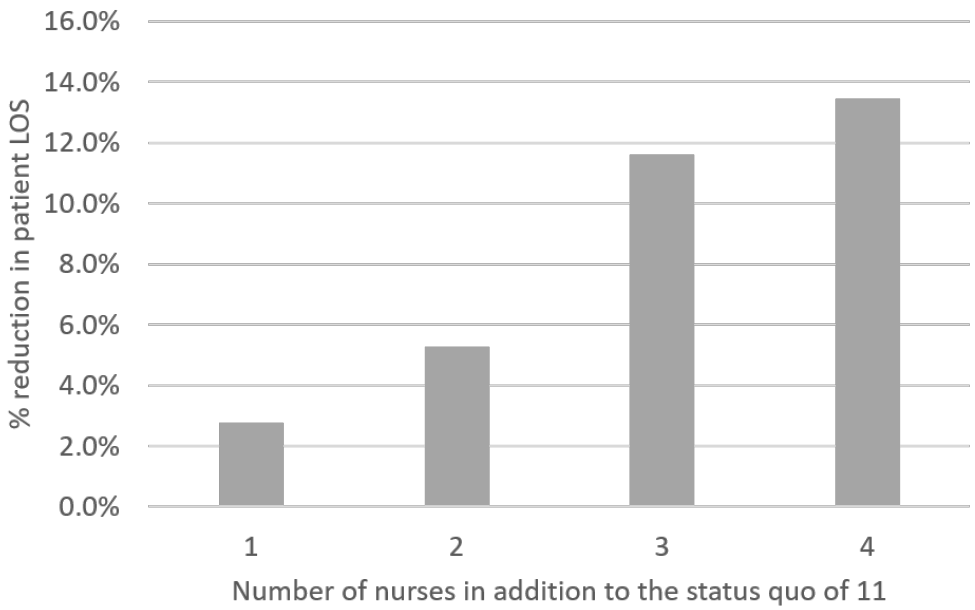


Figure 4.6 Percentage reduction in weighted patient LOS on adding 1, 2, 3, & 4 nurses in addition to 11 considered in the experimental scenario in Section 4.4.2.2.

required threshold. However, care must be taken before setting the operational goals as we noted from the experiment in Table 4.4 that attempting to achieve extreme levels of equality can hurt the efficiency of the system. Our framework thus establishes a decision support system that can provide guidance to a decision-maker such as a lead nurse or hospital administrator on how to staff and route patients depending on the operational characteristics of the system under consideration.

4.5 Conclusions

In this chapter, we developed a framework to obtain optimal staffing, and routing policies for the multi-class many-server pooled queuing system. We used a hybrid method using fluid approximations to queues and simulation to solve the combined routing and staffing problem. We used data from the emergency department of Southeastern Regional Medical Center in Lumberton, NC, to conduct case studies showing the implementation of our framework. Our analyses showed that making small modifications to the routing proportions and staffing policies can lead to an average reduction in patient LOS of up to 23%. Furthermore, we showed that ward workload could also be kept under desired thresholds while noting that attempting to constrain ward threshold too much can lead to an increase in average patient LOS.

We noted during the experimental studies that the choice of operational constraints could lead to vastly different results.

A natural question that arises from an implementation standpoint is about how to use the new routing proportions to send patients to wards. Recall that the status quo routing proportions were obtained from the data retrospectively. Guidance about how to implement the new routing proportions, however, can be established based on the status quo routing proportions obtained from retrospective data. Consider the difference between status quo and optimized routing proportions for patients of severity type 2 in Table 4.2. The status quo suggests that 65% of these patients be sent to critical care, while 35% be sent to minor care. Our optimization, however, modified these values to be 57% and 43%, respectively. The difference between routing 65% versus 57% to critical care is not very clear. From a practical standpoint this difference can only be assessed in terms of aggregate measures across a large number of patients. In other words, the difference between 65% and 57% is that for every 100 patients of severity type 2 entering the ED, 57 of them be sent to critical care as opposed to the 65 being sent to critical care under the status quo. However, such a proportion is often challenging to maintain in the short term as there is sufficient evidence to show the presence of temporal fluctuations in patient arrival rates.

Furthermore, our work assumes that time has no bearing on the nurse workload function. To truly capture workload, it is crucial to incorporate the time since a nurse pool's shift started. This potential temporal variation in nurse workload function is another reason for the difficulty in enforcing routing proportions of patients to wards. Future research could involve developing time-varying proportions that take into account some of these drawbacks.

A second crucial future direction of research could consider the use of transient analysis instead of fluid approximations to analyze the complex queuing models we developed in this Chapter. While fluid approximations are useful in analyzing the average behavior of the system, it does not account for any of the stochastic behavior. Most real systems rarely settle into a steady-state, and the ability to analyze a system in its transient state is often computationally intractable. Furthermore, the results of a transient analysis is a function of the initial conditions of the system, something that a steady-state analysis does not consider. Studying the model developed in this Chapter in its transient state could be a potential future direction of research.

A third important direction for future research stems from consideration for more personalized patient service rates in the queuing model. During the numerical analyses within this chapter, we considered five

patient classes to match the five levels of the Emergency Severity Index (ESI) triage algorithm adopted by Southeastern Regional Medical Center to classify incoming patients. We then inferred functional forms for patient time in service by separating these five severity types depending on the ward they were in, thus leading to 15 possible combinations for the patient time in system functions. However, closer inspection of patient time in service for each of these combinations indicates that a higher level granularity may be possible. Consider the scatter-plot in Fig. 4.3a showing the time in service for patients of severity type 3 receiving care in the minor care ward. Separating these severity type 3 patients into two groups, one requiring higher and the other requiring lower patient times in service, could lead to increased modeling accuracy. Determining the separation threshold (for example, patients requiring more or less than 600 minutes of service) point would require the use of classification algorithms like decision trees trained on data available in the ED such as the primary reason for admission, mode of admission, and initial diagnosis. This framework of increasing the granularity of patient classes within the queuing and fluid models would be better at predicting patient time in service which would then lead to more accurate patient routing and nurse staffing policies.

The final important direction of research relates to understanding the effects of fluid scaling factor η on the optimal policy. We discussed earlier in the chapter that more work is required before ascertaining an appropriate choice for the value of η. While we are able to show from Fig. 4.4 that increasing the value of η leads to a better match between the scaled queuing system and the fluid system, we have not conducted sensitivity analyses about how varying the value of η affects the optimal policies and, subsequently, patient LOS or ward workload. Future research that analyzes the effect of η for queuing systems with non-linear service rates and pooled service could provide insight into the relationship between η and the optimal policy.

Finally, we note that the framework established in this chapter can have applications well beyond the field of healthcare and can benefit any service system that involves customer arrivals into one of several different server pools.

CHAPTER

5

CONCLUSIONS

5.1 Conclusions and a Summary of Contributions

Aside from costing the healthcare industry millions of dollars [Duc14], inefficient resource management has compounding effects on physician and nurse burnout. Employee burnout is one of the biggest challenges facing the healthcare industry [Kam19]. As many as 70% of nurses have reported feeling time pressure, lack of control over work processes, role conflict, and poor relationships between groups and with leadership [Lyn16]. This result of inefficient management is closely associated with staff shortages and high turnover rates. The Bureau of Labor Statistics estimates that 1.2 million vacancies will emerge for registered nurses by 2022 [Usl].

Furthermore, hospitals in the U.S. have seen an average turnover rate of 87.8% [Sol15] of its entire workforce in five years between 2014 and 2019. These side-effects of inefficient management have a coupled relationship: nurse shortages lead to nurses caring for many patients at once, but high workloads can cause burnout leading to staff shortages. A survey commissioned by the Massachusetts Nurses Association in 2018 found that nearly 75% of RNs felt that they were being assigned too many patients at

a time and that they did not have enough time to properly comfort and care for the patients and families as a result. Preventable medical error, another primary concern for hospitals today, is exacerbated by poor nurse-patient ratios. Aiken et al. [Aik02] determined that for every extra nurse, the rate of patient mortality and the rate of failure to rescue decreases by 7%. The use of clinical decision support (CDS) systems has often been considered as a means of battling preventable medical errors [Bat01].

This dissertation introduced a class of models aimed at reducing inefficiencies within hospital management by streamlining patient routing and optimizing nurse staffing. The models touched on three suggested solutions for combating factors negatively impacting hospital standards of care, suggested initially by [Pat11]:

- Use a real-time computerized system to determine whether wards need extra staff.
- Use a group of "floating staff" to assist during busy periods.
- Organize staffing with an eye towards long-term patient health outcomes rather than on addressing immediate patient needs.

There is a growing body of evidence that suggests that unsafe staffing levels lead to high nurse workload and, consequently, adverse patient outcomes. Efficient management of hospital resources is a complex problem. We contend that better management of nursing workloads plays an important role in solving this problem. In all three chapters, we used nurse-patient ratio as a proxy measure for nurse workload within a hospital unit. This one measure, if inefficiently managed, has dire consequences including employee burnout, high turnover rates, staff shortages, and preventable medical errors [Mou17].

The central theme of this work is to improve nurse staffing and patient routing while accounting for nurse workload. In all three chapters, we used nurse-patient ratio as a proxy measure for nurse workload within a hospital unit. The focus of our work has been to demonstrate the building of a framework around any sort of workload measure. Indeed, in Chapter 4 we showed that using nurse-patient ratio as a proxy for workload led to an accurate characterization of patient time in service when compared against real data from SRMC's ED. We the successfully modified patient routing proportions to wards which reduced overall LOS.

An important point that we wish to stress upon the reader is the importance of being able to implement more complicated measures of workload within mathematical frameworks such as the ones developed

in this dissertation. While we used nurse-patient ratio as a proxy for nurse workload throughout this dissertation, advancements in healthcare information technology allows us to leverage a vast amount of data in developing models. One such example is the use of wearable devices to capture physiological measures of workload. Such measures would provide a much more accurate representation of the workload being experienced by ward nurses and would lead to bigger impacts on implementing models like the ones developed in this dissertation. We kept this aspect in mind while defining very general functions of workload across all three chapters. The use of such general functions for workload measures also means added modeling difficulty that we will need to contend with. This was a significant reason for our choice to use numerical solution techniques that rely on computational power as opposed to algorithmic approaches that rely on assumptions that allow for modeling tractability.

5.2 Future Directions

We explore potential directions for future work in this final section. Each proposed extension enhances existing work by making the models more computationally tractable, or by incorporating a new facet about health systems previously assumed absent to allow for ease of modeling or computation.

The first proposed extension is aimed at improving computational performance. In Chapter 3, we developed a patient routing model that considers the workload being experienced by the nurses in the wards, modeled as a function of the nurse-patient ratio in the ward. We currently have a framework to solve for the optimal policy under a 2-ward 1-severity problem setting using complete enumeration or a black box heuristic. In Section 3.5 we discussed how the manipulation of the structure of the LDQBD process and leveraging this unique structure is a means of handling the computational difficulty involved with extending the model to multiple severity types. However, as the problem size grows, enumeration or black box methods start to become computationally intractable. While Gaver's algorithm, detailed in Chapter 3, can solve large problems in a reasonable amount of time, it runs into numeric overflow issues [Gav84] as the state size grows. In order to tackle these computational challenges, we propose an integer programming method using McCormick relaxations to find a lower bound on our minimization problem. Details regarding our proposed formulation using McCormick's relaxation is provided in Appendix D.

Another proposed extension is to adjust our model's objective in Chapters 3 & 4. Both chapters aimed to minimize a patient's LOS. On the one hand, the longer a patient's stay, the higher the risk of

complications. Due to this, the average length of stay is commonly used to gauge the efficiency of a healthcare facility. However, on the other hand, minimizing a patient's stay may be unfavorable if the eventual disposition is to be admitted to the hospital or be discharged to hospice. Patients must not be discharged too quickly and result in hospital readmission. Researchers at VA Medical Centers found that lengths of stay that were longer or shorter than the average length of stay were associated with increased risk of readmission [Kab12]. As a result, the goal of hospitals has been to optimize a patient's length of stay instead of just minimizing it. Related to this, we conducted a preliminary analysis using data from Southeastern Regional Medical Center's (SRMC) emergency department to study the difference in patient LOS based on the final disposition type. We ran statistical tests to confirm our hypothesis that the length of stay for discharged patients is less than the length of stay for admitted patients. First, the two samples have statistically significantly different variances as concluded from the p-value of < 0.01 for an F-test that the ratio of the variance of the two samples is not equal to 1. Besides, we know the average ED LOS for admitted patients is significantly different from the average ED LOS for discharged patients after performing the Welch's t-test (p-value < 0.01). Both results indicate a patient's final disposition type affects their LOS in the ED. We contend that this relationship between patient LOS and final disposition ought to play a role in a hospital's decision making. A proposed direction of future research is *optimizing* a patient's LOS within the modeling frameworks defined in this dissertation instead of *minimizing* patient LOS.

We conclude by noting that engineering advancements aimed towards improving healthcare systems and processes are often fraught with challenges. However, we believe that this dissertation provides insight into challenges surrounding nurse workload and how it affects patient LOS. We hope that future research in health systems engineering considers workload as an important contributing factor when developing models that decide about patient flow and related healthcare processes.

BIBLIOGRAPHY

[Abe73] Abernathy, W. J. et al. "A three-stage manpower planning and scheduling model—a service-sector example". *Operations Research* **21**.3 (1973), pp. 693–711.

[Afe99] Afek, Y. et al. "Convergence complexity of optimistic rate-based flow-control algorithms". *Journal of Algorithms* **30**.1 (1999), pp. 106–143.

[Ago17] Agor, J. et al. "Simulating triage of patients into an internal medicine department to validate the use of an optimization-based workload score". *Proceedings of the 2017 Winter Simulation Conference*. IEEE Press. 2017, p. 234.

[AHR18] AHRQ. *Section 1. The Need to Address Emergency Department Crowding*. 2018.

[Aik02] Aiken, L. H. et al. "Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction". *Jama* **288**.16 (2002), pp. 1987–1993.

[AK09] Al-Kandari, F. & Thomas, D. "Perceived adverse patient outcomes correlated to nurses' workload in medical and surgical wards of selected hospitals in Kuwait". *Journal of Clinical Nursing* **18**.4 (2009), pp. 581–590.

[Ala98] Alanyali, M., Hajek, B., et al. "On large deviations in load sharing networks". *The Annals of Applied Probability* **8**.1 (1998), pp. 67–97.

[Alm13] Almehdawe, E. et al. "A Markovian queueing model for ambulance offload delays". *European Journal of Operational Research* **226**.3 (2013), pp. 602–614.

[And98] Anderson, F. D. et al. "A descriptive, correlational study of patient satisfaction, provider satisfaction, and provider workload at an army medical center". *Military medicine* **163**.2 (1998), pp. 90–94.

[And14] Anderson, R. M. "Stochastic models and data driven simulations for healthcare operations". PhD thesis. Massachusetts Institute of Technology, 2014.

[Ani82] Anick, D. et al. "Stochastic theory of a data-handling system with multiple sources". *Bell System Technical Journal* **61**.8 (1982), pp. 1871–1894.

[Ara11] Ara, R. & Brazier, J. E. "Using health state utility values from the general population to approximate baselines in decision analytic models when condition-specific data are not available". *Value in Health* **14**.4 (2011), pp. 539–545.

[Arm10] Armony, M. & Ward, A. R. "Fair dynamic routing in large-scale heterogeneous-server systems". *Operations Research* **58**.3 (2010), pp. 624–637.

[Arm13] Armony, M. et al. "Critical care in hospitals: When to introduce a step down unit". *Product Oper Manag Google Scholar* (2013).

[Arm15] Armony, M. et al. "On patient flow in hospitals: A data-based queueing-science perspective". *Stochastic Systems* **5**.1 (2015), pp. 146–194.

[Ash17] Ashoo, S. *ED Patient Segmentation*. 2017.

[Ata04] Atar, R. et al. "Scheduling a multi class queue with many exponential servers: Asymptotic optimality in heavy traffic". *The Annals of Applied Probability* **14**.3 (2004), pp. 1084–1134.

[Ata11] Atar, R. et al. "A blind policy for equalizing cumulative idleness". *Queueing Systems* **67**.4 (2011), pp. 275–293.

[Bad09] Badescu, A. L. & Landriault, D. "Applications of fluid flow matrix analytic methods in ruin theory—a review". *RACSAM-Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas* **103**.2 (2009), pp. 353–372.

[Bal18] Ball, J. E. et al. "Post-operative mortality, missed care and nurse staffing in nine countries: A cross-sectional study". *International journal of nursing studies* **78** (2018), pp. 10–15.

[Bar16] Barrett, D. H. et al. *Public health ethics: cases spanning the globe*. Springer, 2016.

[Bar02] Bartal, Y. et al. "Fast, fair and frugal bandwidth allocation in ATM networks". *Algorithmica (New York)* **33**.3 (2002), pp. 272–286.

[Bas15] Bashkin, O. et al. "Organizational Factors affecting length of stay in the emergency department: initial observational study". *Israel journal of health policy research* **4**.1 (2015), p. 38.

[Bat01] Bates, D. W. et al. "Reducing the frequency of errors in medicine using information technology". *Journal of the American Medical Informatics Association* **8**.4 (2001), pp. 299–308.

[Bat13] Bates, K. J. "Floating as a reality: Helping nursing staff keep their heads above water". *Medsurg Nursing* **22**.3 (2013), pp. 197–200.

[Bor99] Bordoloi, S. K. & Weatherby, E. J. "Managerial implications of calculating optimum nurse staffing in medical units". *Health Care Management Review* **24**.4 (1999), pp. 35–44.

[Bor04] Borst, S. et al. "Dimensioning Large Call Centers". *Operations Research* **52**.1 (2004), pp. 17–34.

[Bre05] Brennan, P. F. et al. "Modeling participation in the NHII: operations research approach". *AMIA Annual Symposium Proceedings*. Vol. 2005. American Medical Informatics Association. 2005, p. 76.

[Bre10] Brenner, S. et al. "Modeling and analysis of the emergency department at University of Kentucky Chandler Hospital using simulations". *Journal of emergency nursing* **36**.4 (2010), pp. 303–310.

[Bur91] Burns, L. R. & Wholey, D. R. "The effects of patient, hospital, and physician characteristics on length of stay and mortality". *Medical care* (1991), pp. 251–271.

[But11] Butler, M. et al. "Hospital nurse staffing models and patient and staff-related outcomes". *Cochrane Database of Systematic Reviews* 7 (2011).

[Cao09] Cao, A. et al. “NASA TLX: Software for assessing subjective mental workload”. *Behavior research methods* **41**.1 (2009), pp. 113–117.

[Car08] Carayon, P. & Gurses, A. P. “Nursing workload and patient safety—a human factors engineering perspective” (2008).

[Car09] Carayon, P. & Wood, K. E. “Patient safety”. *Information Knowledge Systems Management* **8**.1-4 (2009), pp. 23–46.

[Car16] Cardoso, T. et al. “Moving towards an equitable long-term care network: A multi-objective and multi-period planning approach”. *Omega* **58** (2016), pp. 69–85.

[Car15] Carnes, K. M. et al. “A cost-benefit analysis of medical scribes and electronic medical record system in an academic urology clinic”. *Urology Practice* **2**.3 (2015), pp. 101–105.

[Cas15] Castro, P. M. “Tightening piecewise McCormick relaxations for bilinear problems”. *Computers & Chemical Engineering* **72** (2015), pp. 300–311.

[Cav02] Cavouras, C. A. “Nurse staffing levels in American hospitals: A 2001 report”. *Journal of Emergency Nursing* **28**.1 (2002), pp. 40–43.

[Che09] Chen, D. et al. “Emergency Materials Distribution Model with Time-Varying Demand and Supply Constraints [J]”. *Logistics Technology* **2** (2009).

[Cle91] Cleveland, W. S. & Grosse, E. “Computational methods for local regression”. *Statistics and computing* **1**.1 (1991), pp. 47–62.

[Coh99] Cohen, M. M. et al. “Nursing workload associated with adverse events in the postanesthesia care unit.” *Anesthesiology* **91**.6 (1999), pp. 1882–1890.

[CC01] Cohen-Charash, Y. & Spector, P. E. “The role of justice in organizations: A meta-analysis”. *Organizational behavior and human decision processes* **86**.2 (2001), pp. 278–321.

[Col01] Colquitt, J. A. “On the dimensionality of organizational justice: A construct validation of a measure.” *Journal of applied psychology* **86**.3 (2001), p. 386.

[Con00] Conn, A. R. et al. *Trust region methods*. Vol. 1. Siam, 2000.

[Con04a] Connelly, L. G. & Bair, A. E. “Discrete event simulation of emergency department activity: A platform for system-level operations research”. *Academic Emergency Medicine* **11**.11 (2004), pp. 1177–1185.

[Con04b] Considine, J. et al. “The Australasian Triage Scale: examining emergency department nurses’ performance using computer and paper scenarios”. *Annals of emergency medicine* **44**.5 (2004), pp. 516–523.

[Cor03] Correia, M. I. T. & Waitzberg, D. L. “The impact of malnutrition on morbidity, mortality, length of hospital stay and costs evaluated through a multivariate model analysis”. *Clinical nutrition* **22**.3 (2003), pp. 235–239.

[Cor05] Corrigan, J. M. et al. “Crossing the quality chasm”. *Building a better delivery system* (2005).

[Cra14] Crawford, F. W. et al. “Estimation for general birth-death processes”. *Journal of the American Statistical Association* **109**.506 (2014), pp. 730–747.

[Dav19] Davis, M. M. et al. “Mailed FIT (fecal immunochemical test), navigation or patient reminders? Using microsimulation to inform selection of interventions to increase colorectal cancer screening in Medicaid enrollees”. *Preventive medicine* **129** (2019), p. 105836.

[Den13] Denton, B. T. “Handbook of healthcare operations management”. *New York: Springer* **10** (2013), pp. 978–1.

[Don08] Donaldson, M. S. “An overview of to err is human: re-emphasizing the message of patient safety”. *Patient safety and quality: An evidence-based handbook for nurses*. Agency for Healthcare Research and Quality (US), 2008.

[Don00] Donaldson, M. S. et al. *To err is human: building a safer health system*. Vol. 6. National Academies Press, 2000.

[Dor16] Doroudi, S. “Stochastic Analysis of Maintenance and Routing Policies in Queueing Systems”. PhD thesis. Carnegie Mellon University, 2016.

[Dor13] Doroudi, S. & Gopalakrishnan, R. “A class of equivalent idle-time-order-based routing policies for heterogeneous multi-server systems”. *arXiv preprint arXiv:1305.6249* (2013).

[Dot09] Dotoli, M. et al. “A continuous Petri net model for the management and design of emergency cardiology departments”. *IFAC Proceedings Volumes* **42**.17 (2009), pp. 50–55.

[Duc14] Duckett, S. J. et al. *Controlling costly care: a billion-dollar hospital opportunity*. Grattan Institute Melbourne, 2014.

[Dud18] Dudin, A. et al. “Analysis of queueing model with processor sharing discipline and customers impatience”. *Operations Research Perspectives* **5** (2018), pp. 245–255.

[Dum85] Dumas, M. B. “Hospital bed utilization: an implemented simulation approach to adjusting and maintaining appropriate levels.” *Health services research* **20**.1 (1985), p. 43.

[DE06] Dziuba-Ellis, J. “Float pools and resource teams: A review of the literature”. *Journal of nursing care quality* **21**.4 (2006), pp. 352–359.

[End06] Endress, A et al. “The critical path method to analyze and modify OR-workflow: Integration of an image documentation system”. *Minimally Invasive Therapy & Allied Technologies* **15**.3 (2006), pp. 177–186.

[Ern04] Ernst, A. T. et al. “An annotated bibliography of personnel scheduling and rostering”. *Annals of Operations Research* **127**.1-4 (2004), pp. 21–144.

[Fan05] Fanjiang, G. et al. *Building a better delivery system: a new engineering/health care partnership*. National Academies Press, 2005.

[Fei77] Feigenson, J. S. et al. “Factors influencing outcome and length of stay in a stroke rehabilitation unit. Part 1. Analysis of 248 unscreened patients–medical and functional prognostic indicators.” *Stroke* **8**.6 (1977), pp. 651–656.

[Fis19] Fishbein, D. et al. “Objective measures of workload in healthcare: a narrative review”. *International Journal of Health Care Quality Assurance* (2019).

[Fri05] Fried, B. et al. *Human resources in healthcare: Managing for success*. Health Administration Press, 2005.

[Gar02] Garnet, O. et al. “Designing a Call Center with Impatient Customers”. *Manufacturing & Service Operations Management* **4**.3 (2002), pp. 208–227.

[Gar00] Garnett, O. & Mandelbaum, A. “An introduction to skills-based routing and its operational complexities”. *Teaching notes* **114** (2000).

[Gav84] Gaver, D. et al. “Finite birth-and-death models in randomly changing environments”. *Advances in applied probability* **16**.4 (1984), pp. 715–731.

[Gla02] Glass, K. P. & Anderson, J. R. “Relative value units: from A to Z (Part I of IV).” *The Journal of medical practice management: MPM* **17**.5 (2002), pp. 225–228.

[Goe14] Goel, S. D. *Textbook of Hospital Administration*. Elsevier Health Sciences, 2014.

[Goo11] Good, E. & Bishop, P. “Willing to walk: A creative strategy to minimize stress related to floating”. *JONA: The Journal of Nursing Administration* **41**.5 (2011), pp. 231–234.

[Gop16] Gopalakrishnan, R. et al. “Routing and staffing when servers are strategic”. *Operations Research* **64**.4 (2016), pp. 1033–1050.

[Gos98] Gosztyla, J & Fowler, S. “Survival skills in the acute care workplace: a" float" pool perspective.” *New Jersey Nurse* **28**.6 (1998), pp. 14–14.

[Gre05] Green, L. V. “Capacity planning and management in hospitals”. *Operations research and health care*. Springer, 2005, pp. 15–41.

[Gre04] Green, L. V. & Kolesar, P. J. “Anniversary article: Improving emergency responsiveness with management science”. *Management Science* **50**.8 (2004), pp. 1001–1014.

[Gre06] Green, L. V. et al. “Managing patient service in a diagnostic medical facility”. *Operations Research* **54**.1 (2006), pp. 11–25.

[Gre13] Green, L. V. et al. ““Nursevendor problem”: Personnel staffing in the presence of endogenous absenteeism”. *Management Science* **59**.10 (2013), pp. 2237–2256.

[Gri17] Griffith, J. D. et al. “Automated dynamic resource allocation for wildfire suppression”. *Lincoln Laboratory Journal* **22**.2 (2017).

[Gri16] Griffiths, P. et al. “Nurse staffing and patient outcomes: Strengths and limitations of the evidence to inform policy and practice. A review and discussion paper based on evidence

reviewed for the National Institute for Health and Care Excellence Safe Staffing guideline development". *International journal of nursing studies* **63** (2016), pp. 213–225.

[Gro08] Gross, D. *Fundamentals of queueing theory*. John Wiley & Sons, 2008.

[Gru06] Gruenberg, D. A. et al. "Factors influencing length of stay in the intensive care unit". *American Journal of critical care* **15**.5 (2006), pp. 502–509.

[Gün10] Günal, M. M. & Pidd, M. "Discrete event simulation for performance modelling in health care: a review of the literature". *Journal of Simulation* **4**.1 (2010), pp. 42–51.

[Gur09] Gurvich, I. & Whitt, W. "Scheduling flexible servers with convex delay costs in many-server service systems". *Manufacturing & Service Operations Management* **11**.2 (2009), pp. 237–253.

[Hal13] Hall, R. "Patient flow". *AMC* **10** (2013), p. 12.

[Hal06] Hall, R. et al. "Modeling patient flows through the healthcare system". *Patient flow: Reducing delay in healthcare delivery*. Springer, 2006, pp. 1–44.

[Hal12] Hall, R. W. et al. *Handbook of healthcare system scheduling*. Springer, 2012.

[Har04a] Haraden, C. & Resar, R. "Patient flow in hospitals: understanding and controlling it better". *Frontiers of health services management* **20**.4 (2004), p. 3.

[Har04b] Harrison, J. M. & Zeevi, A. "Dynamic scheduling of a multiclass queue in the Halfin-Whitt heavy traffic regime". *Operations Research* **52**.2 (2004), pp. 243–257.

[Har05] Harrison, J. M. & Zeevi, A. "A method for staffing large call centers based on stochastic fluid models". *Manufacturing & Service Operations Management* **7**.1 (2005), pp. 20–36.

[Hau00] Hauskrecht, M. & Fraser, H. "Planning treatment of ischemic heart disease with partially observable Markov decision processes". *Artificial Intelligence in Medicine* **18**.3 (2000), pp. 221–244.

[Hee95] Heever, E. Van den. "The use and conservation of indigenous leafy vegetables in South Africa". *workshop 'Genetic Resources of Traditional Vegetables in Africa. Options for Conservation and Use*. 1995, pp. 29–31.

[Hel14] Helm, J. E. & Van Oyen, M. P. "Design and optimization methods for elective hospital admissions". *Operations Research* **62**.6 (2014), pp. 1265–1282.

[Hel80] Helmer, F. T. et al. "Forecasting nursing staffing requirements by intensity-of-care level". *Interfaces* **10**.3 (1980), pp. 50–56.

[Hen5] Hendren, R. *reasons nurses want to leave your hospital. 2011*. 5.

[Her74] Hershey, J. C. et al. "COMPARISON OF NURSE ALLOCATION POLICIES-A MONTE CARLO MODEL". *Decision Sciences* **5**.1 (1974), pp. 58–72.

[Hol11] Holden, R. J. et al. “A human factors framework and study of the effect of nursing workload on patient safety and employee quality of working life”. *BMJ quality & safety* **20**.1 (2011), pp. 15–24.

[Hon15] Honnappa, H. et al. “A queueing model with independent arrivals, and its fluid and diffusion limits”. *Queueing Systems* **80**.1-2 (2015), pp. 71–103.

[Hoo17] Hooey, B. L. et al. “The underpinnings of workload in unmanned vehicle systems”. *IEEE Transactions on Human-Machine Systems* **48**.5 (2017), pp. 452–467.

[Hun07] Hung, G. R. et al. “Computer modeling of patient flow in a pediatric emergency department using discrete event simulation”. *Pediatric emergency care* **23**.1 (2007), pp. 5–10.

[Hur08] Hurst, K. “UK ward design: patient dependency, nursing workload, staffing and quality—an observational study”. *International journal of nursing studies* **45**.3 (2008), pp. 370–381.

[JM03] J Murray, M. “The Canadian Triage and Acuity Scale: A Canadian perspective on emergency department triage”. *Emergency medicine* **15**.1 (2003), pp. 6–10.

[Jun99] Jun, J. et al. “Application of discrete-event simulation in health care clinics: A survey”. *Journal of the operational research society* **50**.2 (1999), pp. 109–123.

[Kab12] Kaboli, P. J. et al. “Associations between reduced hospital length of stay and 30-day readmission rate and mortality: 14-year experience in 129 Veterans Affairs hospitals”. *Annals of internal medicine* **157**.12 (2012), pp. 837–845.

[Kam19] Kamal, A. H. et al. “Prevalence and predictors of burnout among hospice and palliative care clinicians in the US”. *Journal of pain and symptom management* (2019).

[Kan07] Kane, R. L. et al. “Nurse staffing and quality of patient care”. *Evid Rep Technol Assess (Full Rep)* **151**.1 (2007), p. 115.

[Kao81] Kao, E. P. & Tung, G. G. “Aggregate nursing requirement planning in a public health care delivery system”. *Socio-economic planning sciences* **15**.3 (1981), pp. 119–127.

[Kha10] Kharoufeh, J. P. “Level-Dependent Quasi-Birth-and-Death Processes”. *Wiley Encyclopedia of Operations Research and Management Science* (2010).

[Kle67] Kleinrock, L. “Time-shared systems: A theoretical treatment”. *Journal of the ACM (JACM)* **14**.2 (1967), pp. 242–261.

[Kle77] Kleinrock, L. & Kamoun, F. “Hierarchical routing for large networks”. *Computer networks* **1**.3 (1977), pp. 155–174.

[Koç15] Koçağa, Y. L. et al. “Staffing call centers with uncertain arrival rates and co-sourcing”. *Production and Operations Management* **24**.7 (2015), pp. 1101–1117.

[Koo02] Koole, G. & Mandelbaum, A. “Queueing models of call centers: An introduction”. *Annals of Operations Research* **113**.1-4 (2002), pp. 41–59.

[Koo06] Koole, G. & Pot, A. *An overview of routing and staffing algorithms in multi-skill customer contact centers*. 2006.

[KK07] Kopach-Konrad, R. et al. "Applying systems engineering principles in improving health care delivery". *Journal of general internal medicine* **22**.3 (2007), pp. 431–437.

[LF11] Lamy Filho, F. et al. "Staff workload and adverse events during mechanical ventilation in neonatal intensive care units". *Jornal de pediatria* **87**.6 (2011), pp. 487–492.

[Lan04] Lang, T. A. et al. "Nurse–patient ratios: a systematic review on the effects of nurse staffing on patient, nurse employee, and hospital outcomes". *JONA: The Journal of Nursing Administration* **34**.7 (2004), pp. 326–337.

[Lar14] Larrañaga, M. et al. "Index policies for a multi-class queue with convex holding cost and abandonments". *The 2014 ACM international conference on Measurement and modeling of computer systems*. 2014, pp. 125–137.

[Lat18] Latouche, G. & Nguyen, G. "Analysis of fluid flow models". *arXiv preprint arXiv:1802.04355* (2018).

[Lav76] Lave, J. R. & Leinhardt, S. "The cost and length of a hospital stay". *Inquiry* **13**.4 (1976), pp. 327–343.

[Leb15] Lebanik, L. & Britt, S. "Float pool nurses come to the rescue". *Nursing2018* **45**.3 (2015), pp. 50–53.

[Lip75] Lippman, S. A. "Applying a new device in the optimization of exponential queuing systems". *Operations Research* **23**.4 (1975), pp. 687–710.

[Lip98] Lipscomb, J. et al. "Combining expert judgment by hierarchical modeling: An application to physician staffing". *Management Science* **44**.2 (1998), pp. 149–161.

[Liu18] Liu, R. & Xie, X. "Physician Staffing for Emergency Departments with Time-Varying Demand". *INFORMS Journal on Computing* **30**.3 (2018), pp. 588–607.

[Liu10] Liu, Y. & Whitt, W. "A fluid approximation for large-scale service systems". *ACM SIGMETRICS Performance Evaluation Review* **38**.2 (2010), pp. 27–29.

[Liu16] Liu, Z. *Modeling and simulation for healthcare operations management using high performance computing and agent-based model*. Universitat Autònoma de Barcelona, 2016.

[Los89] Loschiavo, J. E. *A Comparison of Nurse Staffing Methods Used by the United States Air Force and Selected Civilian Hospitals*. Tech. rep. Air Force Inst Of Tech Wright-Patterson AFB OH School Of Systems And Logistics, 1989.

[Lyn16] Lyndon, A. "Burnout among health professionals and its effect on patient safety". *Agency of Healthcare Research and Quality* (2016).

[Mae13] Maenhout, B. & Vanhoucke, M. "An integrated nurse staffing and scheduling analysis for longer-term nursing staff allocation problems". *Omega* **41**.2 (2013), pp. 485–499.

[Mag17] Magalhães, A. M. M. d. et al. "Association between workload of the nursing staff and patient safety outcomes". *Revista da Escola de Enfermagem da USP* **51** (2017).

[MR73] Maier-Rothe, C. & Wolfe, H. B. "Cyclical scheduling and allocation of nursing staff". *Socio-Economic Planning Sciences* **7**.5 (1973), pp. 471–487.

[Man95] Mandelbaum, A. & Pats, G. "State-dependent queues: approximations and applications". *Stochastic networks* **71** (1995), pp. 239–282.

[Man04] Mandelbaum, A. & Stolyar, A. L. "Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ-rule". *Operations Research* **52**.6 (2004), pp. 836–855.

[Man12a] Mandelbaum, A. et al. "On fair routing from emergency departments to hospital wards: QED queues with heterogeneous servers". *Management Science* **58**.7 (2012), pp. 1273–1291.

[Man12b] Mandelbaum, A. et al. "On fair routing from emergency departments to hospital wards: QED queues with heterogeneous servers". *Management Science* **58**.7 (2012), pp. 1273–1291.

[MAT] MATLAB. *fmincon Interior Point Algorithm*.

[Maz16] Mazur, L. M. et al. "Toward a better understanding of task demands, workload, and performance during physician-computer interactions". *Journal of the American Medical Informatics Association* **23**.6 (2016), pp. 1113–1120.

[McC76] McCormick, G. P. "Computability of global solutions to factorable nonconvex programs: Part I—Convex underestimating problems". *Mathematical programming* **10**.1 (1976), pp. 147–175.

[McH11] McHugh, M. D. et al. "Nurses' widespread job dissatisfaction, burnout, and frustration with health benefits signal problems for patient care". *Health Affairs* **30**.2 (2011), pp. 202–210.

[McL13] McLay, L. A. & Mayorga, M. E. "A model for optimally dispatching ambulances to emergency calls with classification errors in patient priorities". *IIE Transactions* **45**.1 (2013), pp. 1–24.

[Men14] Mensik, J. "What every nurse should know about staffing". *American Nurse Today* **9**.2 (2014), pp. 1–11.

[Meu92] Meurant, G. "A review on the inverse of symmetric tridiagonal and block tridiagonal matrices". *SIAM Journal on Matrix Analysis and Applications* **13**.3 (1992), pp. 707–728.

[Mil12] Milburn, A. B. "Operations research applications in home healthcare". *Handbook of Healthcare System Scheduling*. Springer, 2012, pp. 281–302.

[Mit12] Mitchell, P. & Golden, R. *Core principles & values of effective team-based health care*. National Academy of Sciences, 2012.

[Mor56] Moran, P. "A probability theory of a dam with a continuous release". *The Quarterly Journal of Mathematics* **7**.1 (1956), pp. 130–137.

[Mos72] Moses, M. “Dispatching and Allocating Servers to Stochastically Failing Networks”. *Management Science* **18**.6 (1972), B–289.

[Mou17] Mousa, A. “Nurse staffing, patient falls and medication errors in Western Australian hospitals: Is there a relationship?” (2017).

[Mue16] Muennig, P. & Bounthavong, M. *Cost-effectiveness analysis in health: a practical approach*. John Wiley & Sons, 2016.

[Naj17] Najman, J. & Mitsos, A. “Tighter McCormick relaxations through subgradient propagation”. *Journal of Global Optimization* (2017), pp. 1–29.

[Cms] *NHE-Fact-Sheet*. 2019.

[Nic18] Nicosia, F. M. et al. “Nurses’ Perspectives on Lean Redesigns to Patient Flow and Inpatient Discharge Process Efficiency”. *Global Qualitative Nursing Research* **5** (2018), p. 2333393618810658. eprint: `https://doi.org/10.1177/2333393618810658`.

[Oet16] Oetelaar, W. van den et al. “Balancing nurses’ workload in hospital wards: study protocol of developing a method to manage workload”. *BMJ open* **6**.11 (2016), e012148.

[Ore11] Oredsson, S. et al. “A systematic review of triage-related interventions to improve patient flow in emergency departments”. *Scandinavian journal of trauma, resuscitation and emergency medicine* **19**.1 (2011), p. 43.

[PL11] Pamela Linzer, M. et al. “What floats a float nurse’s boat?” *Creative nursing* **17**.3 (2011), p. 130.

[Par03] Parachoor, S. B. et al. “Knowledge management system for benchmarking performance indicators using statistical process control (SPC) and Virtual Instrumentation (VI).” *Biomedical sciences instrumentation* **39** (2003), pp. 175–178.

[Par14] Parenti, N. et al. “A systematic review on the validity and reliability of an emergency department triage scale, the Manchester Triage System”. *International journal of nursing studies* **51**.7 (2014), pp. 1062–1069.

[Pat11] Patterson, J. “The effects of nurse to patient ratios”. *Nursing times* **107** (2011), pp. 22–5.

[Pit15] Pitkäaho, T. et al. “Non-linear relationships between nurse staffing and patients’ length of stay in acute care units: Bayesian dependence modelling”. *Journal of advanced nursing* **71**.2 (2015), pp. 458–473.

[Pun06] Punnakitikashem, P. et al. “An optimization-based prototype for nurse assignment”. *Proceedings of the 7th Asian Pacific industrial engineering and management systems conference*. Citeseer. 2006, pp. 17–20.

[Put14] Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[Pya15] Pyati, A. “Float Nursing on the Rise”. *Minority Nurse Magazine* (2015).

[Usl] *Registered Nurses : Occupational Outlook Handbook*. 2019.

[Rub99] Rubenstein, D. et al. "The impact of multicast layering on network fairness". *ACM SIGCOMM Computer Communication Review*. Vol. 29. 4. ACM. 1999, pp. 27–38.

[Rya05] Ryan, J. K. "Systems engineering: Opportunities for health care". *Building a Better Delivery System: A New Engineering/Health Care Partnership* **141** (2005).

[Sag12] Saghafian, S. et al. "Patient streaming as a mechanism for improving responsiveness in emergency departments". *Operations Research* **60**.5 (2012), pp. 1080–1097.

[Sag15] Saghafian, S. et al. "Operations research/management contributions to emergency department patient flow optimization: Review and research prospects". *IIE Transactions on Healthcare Systems Engineering* **5**.2 (2015), pp. 101–123.

[Sau89] Saunders, C. E. et al. "Modeling emergency department operations using advanced computer simulation systems". *Annals of emergency medicine* **18**.2 (1989), pp. 134–140.

[Sch56] Schlager, K. J. "Systems engineering-key to modern development". *IRE Transactions on Engineering Management* 3 (1956), pp. 64–66.

[Sch13] Schmidt, R. et al. "Decision support for hospital bed management using adaptable individual length of stay estimations and shared resources". *BMC medical informatics and decision making* **13**.1 (2013), p. 3.

[Sch15] Schreuders, L. W. et al. "The relationship between nurse staffing and inpatient complications". *Journal of advanced nursing* **71**.4 (2015), pp. 800–812.

[Sim11] Simmons, N. C. & Kuys, S. S. "Trial of an allied health workload allocation model". *Australian Health Review* **35**.2 (2011), pp. 168–175.

[Sim95] Simonian, A. & Guibert, J. "Large deviations approximation for fluid queues fed by a large number of on/off sources". *IEEE Journal on Selected Areas in Communications* **13**.6 (1995), pp. 1017–1027.

[Sir15] Sir, M. Y. et al. "Nurse–patient assignment models considering patient acuity metrics and nurses' perceived workload". *Journal of biomedical informatics* **55** (2015), pp. 237–248.

[Sir17] Sir, M. Y. et al. "Optimization of Multidisciplinary Staffing Improves Patient Experiences at the Mayo Clinic". *Interfaces* **47**.5 (2017), pp. 425–441.

[Sis07] Sisselman, M. E. & Whitt, W. "Value-based routing and preference-based routing in customer contact centers". *Production and Operations Management* **16**.3 (2007), pp. 277–291.

[Sol15] Solutions, N. N. et al. "2019 National Healthcare Retention & RN Staffing Report" (2015).

[Sou00] Souder, E. & O'Sullivan, P. S. "Nursing documentation versus standardized assessment of cognitive status in hospitalized medical patients". *Applied Nursing Research* **13**.1 (2000), pp. 29–36.

[Sta04] Stanton, M. W. *Hospital nurse staffing and quality of care*. Agency for Healthcare Research and Quality Rockville, MD, 2004.

[Ste06] Steenstra, I. A. et al. "Economic evaluation of a multi-stage return to work program for workers on sick-leave due to low back pain". *Journal of occupational rehabilitation* **16**.4 (2006), pp. 557–578.

[Sto15] Stolldorf, D. P. & Jones, C. B. "Deployment of rapid response teams by 31 hospitals in a statewide collaborative". *Joint Commission journal on quality and patient safety* **41**.4 (2015), AP1–AP3.

[Swa19] Swan, B. et al. "Evaluating an Emergency Department Care Redesign: A Simulation Approach". *2019 Winter Simulation Conference (WSC)*. 2019, pp. 1137–1147.

[Tan04] Tanabe, P. et al. "The Emergency Severity Index (version 3) 5-level triage system scores predict ED resource consumption". *Journal of Emergency Nursing* **30**.1 (2004), pp. 22–29.

[Tan03] Tanenbaum, A. S. & service), S. T. B. O. O. *Computer networks*. Prentice Hall Professional, 2003.

[Tho20] Thore, C.-J. "Fminsdp–a code for solving optimization problems with matrix inequality constraints". *MATLAB Cent. File Exch.* (2020).

[Tho09] Thorwarth, M. & Arisha, A. "Application of discrete-event simulation in health care: a review" (2009).

[Thu07] Thungjaroenkul, P. et al. "The impact of nurse staffing on hospital costs and patient length of stay: a systematic review". *Nursing Economics* **25**.5 (2007), p. 255.

[Tri76] Trivedi, V. M. & Warner, D. M. "A branch and bound algorithm for optimum allocation of float nurses". *Management Science* **22**.9 (1976), pp. 972–981.

[Tse09] Tseytlin, Y. "Queueing Systems with Heterogeneous Servers: On Fair Routing of Patients in Emergency Departments. Technion M. Sc". PhD thesis. Thesis, April 2009. Available at `http://ie.technion.ac.il/serveng/References/thesis-yulia.pdf`. 6, 7, 2009.

[Tur86] Turner, J. "New directions in communications(or which way to the information age?)" *IEEE communications Magazine* **24**.10 (1986), pp. 8–15.

[Tuy98] Tuy, H. et al. *Convex analysis and global optimization*. Springer, 1998.

[Twi09] Twigg, D. & Duffield, C. "A review of workload measures: a context for a new staffing methodology in Western Australia". *International journal of nursing studies* **46**.1 (2009), pp. 132–140.

[Twi11] Twigg, D. et al. "The impact of the nursing hours per patient day (NHPPD) staffing method on patient outcomes: a retrospective analysis of patient and staffing data". *International journal of nursing studies* **48**.5 (2011), pp. 540–548.

[Upe07] Upenieks, V. V. et al. “Assessing nursing staffing ratios: variability in workload intensity”. *Policy, Politics, & Nursing Practice* **8**.1 (2007), pp. 7–19.

[Vér11] Véricourt, F. d. & Jennings, O. B. “Nurse staffing in medical units: A queueing perspective”. *Operations Research* **59**.6 (2011), pp. 1320–1331.

[Ver09] Vermeulen, I. B. et al. “Adaptive resource allocation for efficient patient scheduling”. *Artificial intelligence in medicine* **46**.1 (2009), pp. 67–80.

[Wal04] Wallace, R. B. & Whitt, W. “Resource pooling and staffing in call centers with skill-based routing”. *Operations Research* **7**.4 (2004), pp. 276–294.

[Wal05] Wallace, R. B. & Whitt, W. “A staffing algorithm for call centers with skill-based routing”. *Manufacturing & Service Operations Management* **7**.4 (2005), pp. 276–294.

[Wan12] Wang, J. et al. “Reducing length of stay in emergency department: A simulation study at a community hospital”. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* **42**.6 (2012), pp. 1314–1322.

[Wan14] Wang, W.-Y. & Gupta, D. “Nurse absenteeism and staffing strategies for hospital inpatient units”. *Manufacturing & Service Operations Management* **16**.3 (2014), pp. 439–454.

[War13] Ward, A. R. & Armony, M. “Blind fair routing in large-scale service systems with heterogeneous customers and servers”. *Operations Research* **61**.1 (2013), pp. 228–243.

[War76] Warner, D. M. “Scheduling nursing personnel according to nursing preference: A mathematical programming approach”. *Operations Research* **24**.5 (1976), pp. 842–856.

[Wei07] Weissman, J. S. et al. “Hospital workload and adverse events”. *Med. Care* **45**.5 (2007), pp. 448–455.

[Wel09] Welch, S. J. “Patient segmentation: Redesigning flow”. *Emergency Medicine News* **31**.8 (2009).

[Wen99] Weng, M. L. & Houshmand, A. A. “Healthcare simulation: a case study at a local clinic”. *Proceedings of the 31st conference on Winter simulation: Simulation—a bridge to the future-Volume 2*. 1999, pp. 1577–1584.

[Whi10] Whitehead, S. J. & Ali, S. “Health outcomes in economic evaluation: the QALY and utilities”. *British medical bulletin* **96**.1 (2010), pp. 5–21.

[Whi04] Whitt, W. “Efficiency-driven heavy-traffic approximations for many-server queues with abandonments”. *Management Science* **50**.10 (2004), pp. 1449–1461.

[Whi06a] Whitt, W. “A multi-class fluid model for a contact center with skill-based routing”. *AEU-International Journal of Electronics and Communications* **60**.2 (2006), pp. 95–102.

[Whi06b] Whitt, W. “Fluid models for multiserver queues with abandonments”. *Operations research* **54**.1 (2006), pp. 37–54.

[Wil11] Wiler, J. L. et al. “Review of modeling approaches for emergency department patient flow and crowding research”. *Academic Emergency Medicine* **18**.12 (2011), pp. 1371–1379.

[Wol16] Wolowacz, S. E. et al. “Estimating health-state utility for economic models in clinical studies: an ISPOR good research practices task force report”. *Value in Health* **19**.6 (2016), pp. 704–719.

[Wri06] Wright, P. D. et al. “Reexamining the nurse scheduling problem: Staffing ratios and nursing shortages”. *Decision Sciences* **37**.1 (2006), pp. 39–70.

[YT14] Yom-Tov, G. B. & Mandelbaum, A. “Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing”. *Manufacturing & Service Operations Management* **16**.2 (2014), pp. 283–299.

[Yoo03] Yoon, P. et al. “Analysis of factors influencing length of stay in the emergency department”. *Canadian Journal of Emergency Medicine* **5**.3 (2003), pp. 155–161.

[You19] Yousefi, N. et al. “Appointment scheduling model in healthcare using clustering algorithms”. *arXiv preprint arXiv:1905.03083* (2019).

[Zel11] Zeltyn, S. et al. “Simulation-based models of emergency departments: Operational, tactical, and strategic staffing”. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* **21**.4 (2011), pp. 1–25.

[Zem70] Zemach, R. “A model of health-service utilization and resource allocation”. *Operations Research* **18**.6 (1970), pp. 1071–1086.

[Zim12] Zimring, C. & Seo, H.-B. “Making acuity-adaptable units work: Lessons from the field”. *HERD: Health Environments Research & Design Journal* **5**.3 (2012), pp. 115–128.

APPENDICES

APPENDIX

A

RESULTS SHOWING VARIATION IN STRUCTURE OF ODP ACCORDING TO EXPERIMENTS IN CHAPTER 2

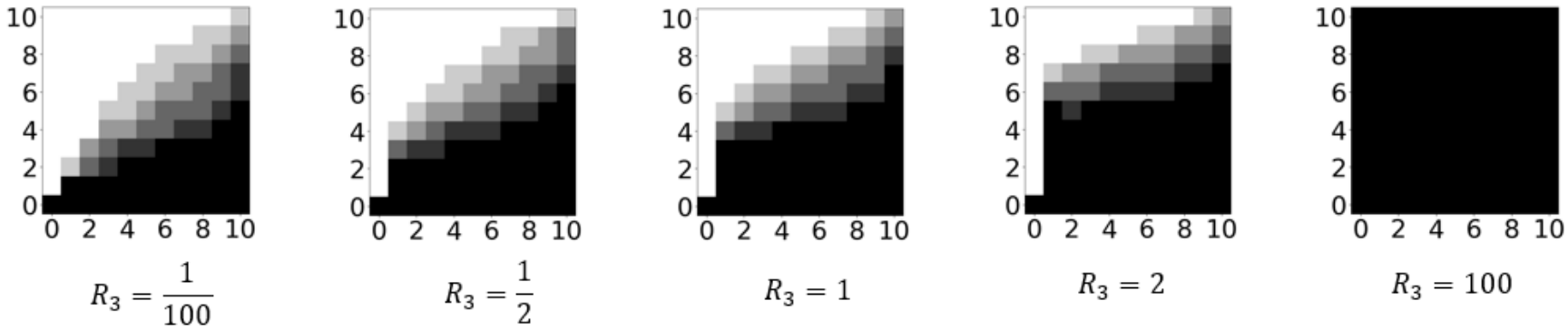


Figure A.1 Variation in structure of ODP when problem parameters outlined in Table 2.1 are modified to increase (by multiplier R_3) the coefficient for the Health State Utility function for the less severe ward. The axes represent the number of patients in wards 1 (x-axis) and 2 (y-axis) while the colorbar represents the number of floating nurses that must be assigned to ward 1.

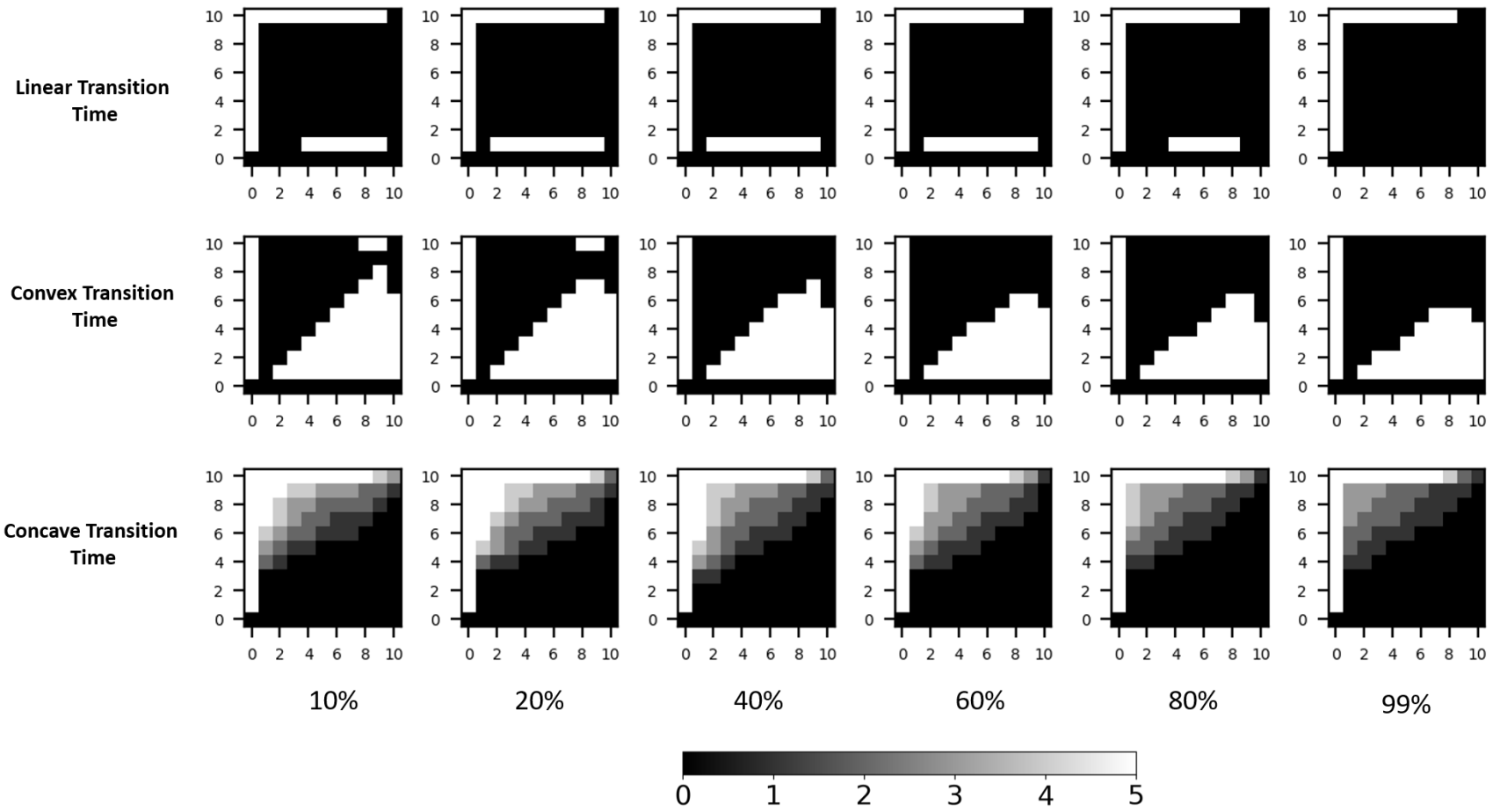


Figure A.2 ODP for modified objective function considering only one-time cost for patient discharge to hospice under varying values of $\hat{\rho}$. The colorbar shows the number of floating resources being assigned to the less severe ward.

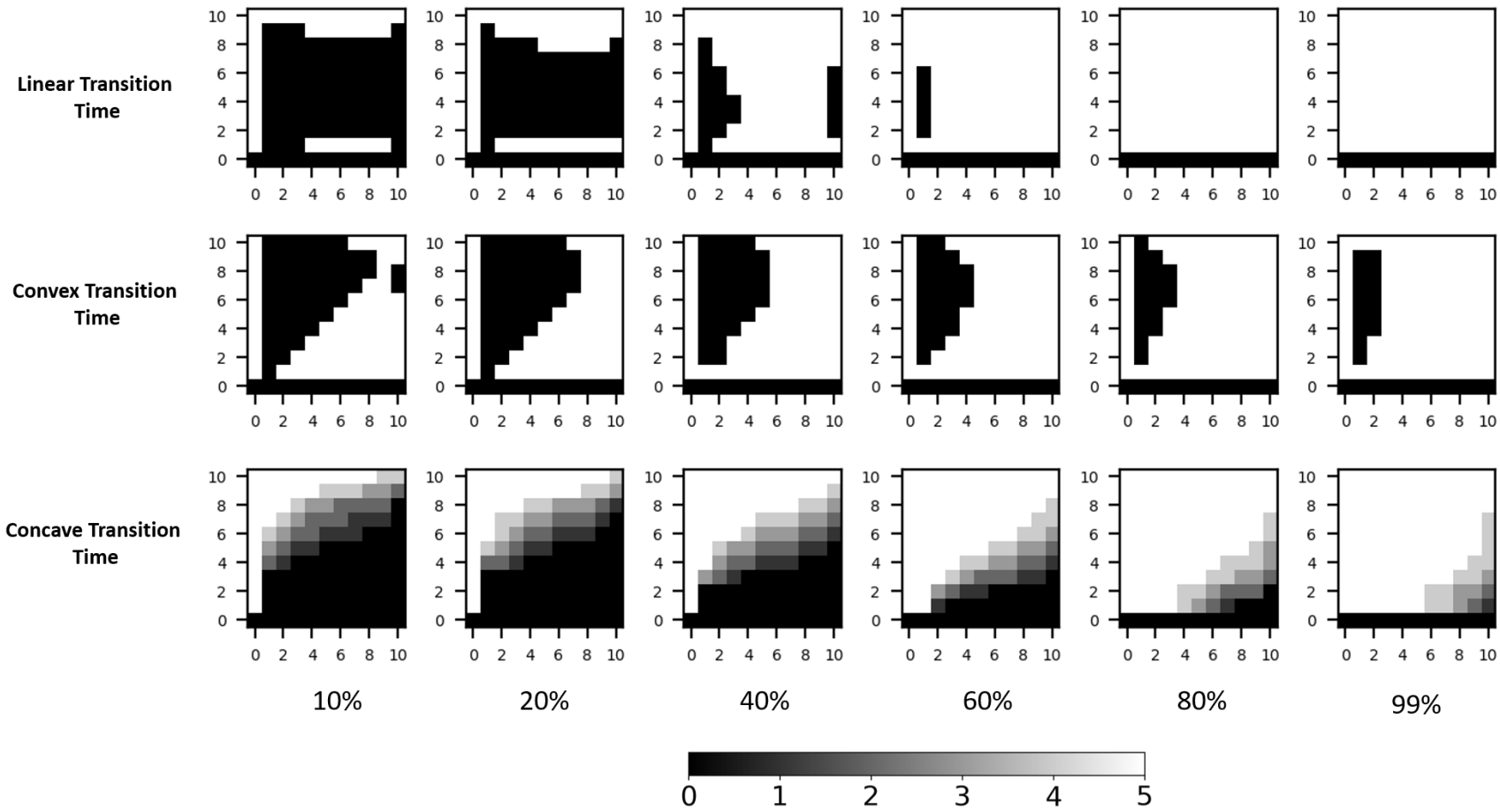


Figure A.3 ODP for modified objective function considering only one-time reward for patient recovery under varying values of $\hat{\rho}$. The colorbar shows the number of floating resources being assigned to the less severe ward.

A.1 Results from experimenting with various heuristics and comparing them against optimal dynamic policy as outlined in Section 2.5

Table A.1.1 A comparison of the heuristics' performance against the ODP on varying the multiplier corresponding to the entire HSU function.

R_2	**ODP**	**AN**		**PA**		**OSP**	
	g^*	g^H	Gap^H	g^H	Gap^H	g^H	Gap^H
1/100	78.84	78.64	0.3%	77.79	1.3%	78.54	0.4%
1/2	98.53	98.31	0.2%	97.53	1.0%	98.15	0.4%
1	181.62	181.15	0.25%	180.62	0.55%	180.72	0.5%
2	601.41	595.34	1.0%	596.07	0.9%	596.85	0.8%
100	10,777	10,432.6	3.2%	10,463.17	2.9%	10,702.43	0.7%

Table A.1.2 A comparison of the heuristics' performance against the ODP on varying the multiplier corresponding to HSU function coefficient for the less severe ward.

R_2	**ODP**	**AN**		**PA**		**OSP**	
	g^*	g^H	Gap^H	g^H	Gap^H	g^H	Gap^H
1/100	80.90	80.62	0.34%	79.64	1.55%	80.50	0.49%
1/2	132.14	131.83	0.23%	130.87	0.96%	131.59	0.42%
1	181.62	181.15	0.25%	180.62	0.55%	180.72	0.49%
2	289.28	288.60	0.23%	287.71	0.54%	287.99	0.45%
100	10693.39	10530.87	1.52%	10534.28	1.49%	10624.47	0.64%

Table A.1.3 A comparison of the performance of heuristics against the optimal policy when the objective function only considers the one-time reward of patient discharge home.

	Linear				Convex				Concave			
	ODP	**AN**	**PA**	**OSP**	**ODP**	**AN**	**PA**	**OSP**	**ODP**	**AN**	**PA**	**OSP**
10%	45.78	45.04 (1.62%)	45.15 (1.38%)	44.16 (3.53%)	43.73	42.34 (3.18%)	42.06 (3.81%)	41.85 (4.31%)	47.76	47.28 (1.01%)	47.64 (0.25%)	46.87 (1.87%)
20%	79.63	78.36 (1.59%)	78.56 (1.34%)	78.66 (1.21%)	75.67	72.76 (3.84%)	72.66 (3.97%)	74.66 (1.33%)	85.48	84.86 (0.73%)	85.44 (0.05%)	84.82 (0.78%)
40%	127.60	127.14 (0.36%)	127.26 (0.27%)	127.16 (0.34%)	122.71	120.71 (1.63%)	120.72 (1.62%)	122.50 (0.17%)	139.65	139.47 (0.13%)	139.57 (0.06%)	139.33 (0.23%)
60%	149.27	149.11 (0.11%)	148.84 (0.29%)	149.26 (0.01%)	144.01	143.43 (0.41%)	143.37 (0.45%)	143.90 (0.08%)	164.09	163.97 (0.07%)	163.66 (0.26%)	164.05 (0.03%)
80%	152.99	152.78 (0.14%)	152.22 (0.50%)	152.99 (0.00%)	147.93	147.77 (0.10%)	147.68 (0.17%)	147.83 (0.07%)	168.34	168.02 (0.19%)	166.95 (0.82%)	168.33 (0.00%)
99%	149.77	149.52 (0.17%)	148.82 (0.64%)	149.77 (0.00%)	145.06	144.99 (0.05%)	144.89 (0.11%)	145.04 (0.02%)	164.81	164.33 (0.29%)	162.82 (1.21%)	164.80 (0.00%)

Table A.1.4 A comparison of the performance of heuristics against the optimal policy when the objective function only considers the one-time cost of patient discharge to hospice. We note here that all values in this table are negative since the one-time cost of patient discharge to hospice is a negative value.

	Linear				Convex				Concave			
	ODP	**AN**	**PA**	**OSP**	**ODP**	**AN**	**PA**	**OSP**	**ODP**	**AN**	**PA**	**OSP**
10%	-12.82	-13.56 (5.80%)	-13.55 (5.72%)	-14.53 (13.40%)	-14.97	-16.36 (9.31%)	-16.64 (11.19%)	-16.87 (12.70%)	-10.29	-10.77 (4.67%)	-10.67 (3.71%)	-11.51 (11.92%)
20%	-30.99	-32.29 (4.20%)	-32.26 (4.13%)	-32.34 (4.36%)	-35.14	-38.07 (8.35%)	-38.18 (8.65%)	-36.17 (2.94%)	-24.19	-24.83 (2.64%)	-24.71 (2.16%)	-25.37 (4.89%)
40%	-64.85	-66.33 (2.28%)	-66.29 (2.21%)	-65.47 (0.95%)	-70.31	-73.00 (3.82%)	-72.96 (3.76%)	-70.53 (0.30%)	-52.55	-53.14 (1.12%)	-53.03 (0.93%)	-53.34 (1.52%)
60%	-88.54	-89.97 (1.62%)	-89.93 (1.56%)	-89.17 (0.71%)	-94.08	-95.79 (1.82%)	-95.75 (1.78%)	-94.14 (0.07%)	-74.50	-75.31 (1.08%)	-75.15 (0.88%)	-75.38 (1.18%)
80%	-103.51	-104.77 (1.22%)	-104.75 (1.20%)	-104.22 (0.69%)	-108.63	-109.68 (0.96%)	-109.67 (0.95%)	-108.67 (0.03%)	-89.47	-90.55 (1.21%)	-90.34 (0.97%)	-90.55 (1.21%)
99%	-111.97	-113.11 (1.02%)	-113.10 (1.01%)	-112.66 (0.62%)	-116.62	-117.32 (0.60%)	-117.32 (0.60%)	-116.64 (0.02%)	-98.58	-99.92 (1.35%)	-99.60 (1.03%)	-99.79 (1.23%)

APPENDIX

B

GENERATOR MATRICES FOR LDQBD PROCESS IN CHAPTER 3

This appendix gives the generator matrices for the LDQBD process outlined in Section 3 of Chapter 3. Recall that our state space set was defined as

$$\begin{aligned}\{&(0,0,0),(0,1,0),(0,2,0),...,(0,M,0),\\ &(1,0,0),(1,1,0),(1,2,0),..,(1,M,0),\\ &...,\\ &(M,0,0),(M,1,0),(M,2,0),(M,M,0),(M,M,1),(0,M,2),...,(M,M,W)\end{aligned} \tag{B.1}$$

Our LDQBD process is defined using phases and levels such that the process can only go up or down one level at a time. Within each level is a set of phases that do not have this same restriction. In the context of our problem, we define our levels as the number of patients in the first unit and the phases within each level as the number of patients in the second unit. Each of the $M+1$ levels has $M+1$ phases except for

the last level which has $M+1+W$ phases to represent the number of waiting patients. The matrix $Q(\bar{\beta})$ can be re-written via levels $L_0, L_1, .., L_M$ as

$$
Q(\bar{\beta}) = \begin{array}{c} \\ L_0 \\ L_1 \\ L_2 \\ \vdots \\ L_M \\ L_{M-1} \end{array}
\begin{array}{c}
\begin{array}{ccccccc} L_0 & L_1 & L_2 & \dots & L_M & L_{M-1} \end{array} \\
\begin{pmatrix}
A_1(0,\bar{\beta}) & A_0(0,\bar{\beta}) & \dots & & & \\
A_2(1,\bar{\beta}) & A_1(1,\bar{\beta}) & A_0(1,\bar{\beta}) & & & \\
 & A_2(2,\bar{\beta}) & A_1(2,\bar{\beta}) & & & \vdots \\
\vdots & & & \ddots & & \\
 & & & \dots & A_1(M-1,\bar{\beta}) & A_0(M-1,\bar{\beta}) \\
 & & & \dots & A_1(M,\bar{\beta}) & A_0(M,\bar{\beta})
\end{pmatrix}
\end{array}
$$

$A_0(i,\bar{\beta})$, $A_1(i,\bar{\beta})$ and $A_2(i,\bar{\beta})$ are the generator matrices for the QBD process, where i gives the level (number of patients in the first unit). Below we outline the structure for each of $A_0(i,\bar{\beta})$, $A_1(i,\bar{\beta})$ and $A_2(i,\bar{\beta})$.

B.1 When $i = 1,2,3,..,M-1$

$$
A_2(i,\bar{\beta}) = \begin{array}{c} \\ (i,0) \\ (i,1) \\ (i,2) \\ \vdots \\ (i,M-1) \\ (i,M) \end{array}
\begin{array}{c}
\begin{array}{ccccccc} (i-1,0) & (i-1,1) & (i-1,2) & \dots & (i-1,M-1) & (i-1,M) \end{array} \\
\begin{pmatrix}
i\mu_1(i) & & \dots & & & \\
 & i\mu_1(i) & & & & \\
 & & i\mu_1(i) & & & \vdots \\
\vdots & & & \ddots & & \\
 & & & \dots & i\mu_1(i) & \\
 & & & \dots & & i\mu_1(i)
\end{pmatrix}
\end{array}
$$

$$A_1(i,\bar{\beta}) = \begin{array}{c} \\ (i,0) \\ (i,1) \\ (i,2) \\ \vdots \\ (i,M-1) \\ (i,M) \end{array} \begin{array}{c} \begin{array}{ccccccc} (i,0) & (i,1) & (i,2) & \dots & (i,M-1) & (i,M) \end{array} \\ \begin{pmatrix} -\lambda & P_2\lambda & & \dots & & \\ \mu_2(1) & -\lambda & P_2\lambda & & & \\ & 2\mu_2(2) & -\lambda & & & \vdots \\ \vdots & & & \ddots & & \\ & & & \dots & -\lambda & P_2\lambda \\ & & & \dots & M\mu_2(M) & -\lambda \end{pmatrix} \end{array}$$

$$A_0(i,\bar{\beta}) = \begin{array}{c} \\ (i,0) \\ (i,1) \\ (i,2) \\ \vdots \\ (i,M-1) \\ (i,M) \end{array} \begin{array}{c} \begin{array}{ccccccc} (i+1,0) & (i+1,1) & (i+1,2) & \dots & (i+1,M-1) & (i+,M-1) \end{array} \\ \begin{pmatrix} P_1\lambda & & & \dots & & \\ & P_1\lambda & & & & \\ & & P_1\lambda & & & \vdots \\ \vdots & & & \ddots & & \\ & & & \dots & P_1\lambda & \\ & & & \dots & & P_1\lambda \end{pmatrix} \end{array}$$

B.2 When $i = M$

$$A_2(i,\bar{\beta}) = \begin{array}{c} \\ (i,0) \\ (i,1) \\ (i,2) \\ \vdots \\ (i,M-1) \\ (i,M) \\ (i,M,1) \\ \vdots \\ (i,M,W) \end{array} \begin{array}{c} \begin{array}{ccccccc} (M-1,0) & (M-1,1) & (M-1,2) & \dots & (M-1,M-1) & (M-1,M) \end{array} \\ \begin{pmatrix} i\mu_1(i) & & & \dots & & \\ & i\mu_1(i) & & & & \\ & & i\mu_1(i) & & & \vdots \\ \vdots & & & \ddots & & \\ & & & \dots & i\mu_1(i) & \\ & & & \dots & & i\mu_1(i) \\ & & & & & \\ & & & \vdots & & \end{pmatrix} \end{array}$$

$$
A_1(i,\bar{\beta}) \;=\;
\begin{array}{c}
\\
(M,0)\\
\\
(M,1)\\
\\
(M,2)\\
\\
\vdots\\
\\
(M,M-1)\\
\\
(M,M)\\
\\
(M,M,1)\\
\\
\vdots\\
\\
(M,M,W)
\end{array}
\begin{array}{c}
\begin{array}{ccccccccc}
(M,0) & (M,1) & (M,2) & \dots & (M,M-1) & (M,M) & \dots & (M,M,W)
\end{array}\\
\left(\begin{array}{ccccccccc}
-P_2\lambda & & & & & & & \\
-\mu_2(1) & P_2\lambda & \dots & & & & & \\
 & -P_2\lambda & & & & & & \\
\mu_2(1) & -M\mu_1(M) & P_2\lambda & & & & & \\
 & -\mu_2(1) & & & & & & \\
 & & -P_2\lambda & & & & & \\
 & 2\mu_2(2) & -M\mu_1(M) & & \vdots & & & \\
 & & -2\mu_2(2) & & & & & \\
\vdots & & & \ddots & & & & \\
 & & & & -P_2\lambda & & & \\
 & & & \dots & -M\mu_1(M) & P_2\lambda & & \\
 & & & & -(M-1)\mu_2(M-1) & & & \\
 & & & & & -\lambda & & \\
 & & & \dots & M\mu_2(M) & -M(\mu_1(M) & & \\
 & & & & & +\mu_2(M)) & & \\
 & & & & & M(\mu_1(M) & & \\
 & & & & & +\mu_2(M)) & & \\
 & & & & & & \ddots & \\
 & & & & & & & -M(\mu_1(M) \\
 & & & & & & & +\mu_2(M))
\end{array}\right)
\end{array}
$$

APPENDIX

C

COMPLETE FORMULATION OF OPTIMIZATION PROBLEM IN CHAPTER 3

$$\underset{\bar{\alpha},\bar{\beta}}{\textbf{minimize}} \quad \frac{\sum_{i,j,k}[i+j+k]\alpha_{i,j,k}}{\lambda[1-\alpha_{M,M,W}]}$$

subject to

$$\alpha_{i,j,k}[j\mu_2(j,c_2)+i\mu_1(i,c_i)+\lambda] = [i+1]\mu_1(i+1,c_1)\alpha_{i+1,j,k}$$
$$+[j+1]\mu_2(j+1,c_2)\alpha_{i,j+1,k}+\lambda P_1(i-1,j,\bar{\beta})\alpha_{i-1,j,k}+\lambda P_2(i,j-1,\bar{\beta})\alpha_{i,j-1,k},$$
$$i<M, j<M, k=0$$

$$\alpha_{i,j,k}[j\mu_2(j,c_2)+i\mu_1(i,c_1)+\lambda] =$$
$$+[j+1]\mu_2(j+1,c_2)\alpha_{i,j+1,k}+\lambda P_1(i-1,j,\bar{\beta})\alpha_{i-1,j,k}+\lambda P_2(i,j-1,\bar{\beta})\alpha_{i,j-1,k},$$
$$i=M, j<M, k=0$$

$$\alpha_{i,j,k}[j\mu_2(j,c_2)+i\mu_1(i,c_1)+\lambda] =$$
$$+[i+1]\mu_1(i+1,c_1)\alpha_{i+1,j,k}+\lambda P_1(i-1,j,\bar{\beta})\alpha_{i-1,j,k}+\lambda P_2(i,j-1,\bar{\beta})\alpha_{i,j-1,k},$$
$$i<M, j=M, k=0$$

$$\alpha_{i,j,k}[j\mu_2(j,c_2)+i\mu_1(i,c_1)+\lambda] = i\mu_1(i,c_1)\alpha_{i,j,k+1}+j\mu_2(j,c_2)\alpha_{i,j,k+1}+\lambda[\alpha_{i-1,j,k}+\alpha_{i,j-1,k}]$$
$$i=M, j=M, k=0$$

$$\alpha_{i,j,k}[j\mu_2(j,c_2)+i\mu_1(i,c_1)+\lambda] = i\mu_1(i,c_1)\alpha_{i,j,k+1}+j\mu_2(j,c_2)\alpha_{i,j,k+1}+\lambda\alpha_{i,j,k-1},$$
$$i=M, j=M, 0<k<W$$

$$\sum_{i,j,k}\alpha_{i,j,k}=1$$
$$\Psi(\Gamma_1,\Gamma_2)\in\psi,$$
$$0\le\alpha_{i,j,k}\le 1, \qquad i\le M, j\le M, k\le W$$
$$-\infty\le\beta_1,\beta_2\le 0$$

(C.1)

APPENDIX

D

MCCORMICK'S RELAXATION TO SOLVE MULTI-SEVERITY PATIENT ROUTING PROBLEM IN CHAPTER 3

In Chapter 3 we developed a patient routing model that assigns incoming patients to wards in the emergency department based on the workload being experienced by the nurses in the wards. We model the workload of the nurses as a function of the nurse-patient ratio in the ward. We currently have a framework to solve for the optimal policy under a 2-ward 1-severity problem setting using complete enumeration or a black box heuristic. In Section 3.5 we discussed a means of handling the computational difficulty involved with extending the model to multiple severity types by manipulating the structure of the LDQBD process and leveraging its special structure. However, as the problem size begins to grow, enumeration/black box methods start to become computationally intractable. Additionally, while Gaver's algorithm can solve large problems in a reasonable amount of time, it runs into numeric overflow issues [Gav84] as the state size grows. In order to tackle these computational challenges, we propose an integer

programming method using McCormick relaxations to find a lower bound on our minimization problem.

In order to convert the optimization problem given in Section 3.3.2 into a linear programming problem, we first linearize the objective function. Once this is done, we are able to show that when the workload function is assumed to be linear, the constraints of the optimization problem may be -rewritten as a bilinear equality constraints. We then use a relaxation technique to convert the bi-linear equality constraints into linear equality constraints.

Our first step is to linearize the objective function. To do this we first define a new variable δ to replace $\frac{1}{1-\alpha_{M,M,W}}$. We then multiply every equation that doesn't feature $\alpha_{M,M,W}$ by δ. Next, performing some simple algebraic manipulations of the equations that feature the term $\alpha_{M,M,W}$ and replacing the term $\alpha_{i,j,k}\delta$ throughout the model to a new term $\alpha^*_{i,j,k}$ allows us to reformulate the problem in the following manner.

$$\underset{\bar{\alpha}^*,\bar{\beta}}{\text{minimize}} \quad \frac{\sum_{i,j,k}[i+j+k]\alpha^*_{i,j,k}}{\lambda}$$

subject to

$$\alpha^*_{i,j,k}[j\mu_2(j,c_2)+i\mu_1(i,c_i)+\lambda] = [i+1]\mu_1(i+1,c_1)\alpha^*_{i+1,j,k}$$
$$+[j+1]\mu_2(j+1,c_2)\alpha^*_{i,j+1,k}+\lambda P_1(i-1,j,\bar{\beta})\alpha^*_{i-1,j,k}+\lambda P_2(i,j-1,\bar{\beta})\alpha^*_{i,j-1,k},$$
$$i<M, j<M, k=0$$

$$\alpha^*_{i,j,k}[j\mu_2(j,c_2)+i\mu_1(i,c_1)+\lambda] =$$
$$+[j+1]\mu_2(j+1,c_2)\alpha^*_{i,j+1,k}+\lambda P_1(i-1,j,\bar{\beta})\alpha^*_{i-1,j,k}+\lambda P_2(i,j-1,\bar{\beta})\alpha^*_{i,j-1,k},$$
$$i=M, j<M, k=0$$

$$\alpha^*_{i,j,k}[j\mu_2(j,c_2)+i\mu_1(i,c_1)+\lambda] =$$
$$+[i+1]\mu_1(i+1,c_1)\alpha^*_{i+1,j,k}+\lambda P_1(i-1,j,\bar{\beta})\alpha^*_{i-1,j,k}+\lambda P_2(i,j-1,\bar{\beta})\alpha^*_{i,j-1,k},$$
$$i<M, j=M, k=0$$

$$\alpha^*_{i,j,k}[j\mu_2(j,c_2)+i\mu_1(i,c_1)+\lambda] = i\mu_1(i,c_1)\alpha^*_{i,j,k+1}$$
$$+j\mu_2(j,c_2)\alpha^*_{i,j,k+1}+\lambda[\alpha^*_{i-1,j,k}+\alpha^*_{i,j-1,k}],$$
$$i=M, j=M, k=0$$

$$\alpha^*_{i,j,k}[j\mu_2(j,c_2)+i\mu_1(i,c_1)+\lambda] = i\mu_1(i,c_1)\alpha^*_{i,j,k+1}$$
$$+j\mu_2(j,c_2)\alpha^*_{i,j,k+1}+\lambda\alpha^*_{i,j,k-1}, 0<k<W$$
$$i=M, j=M,$$

$$\sum_{(i,j,k)\neq(M,M,W)} \alpha^*_{i,j,k} = 1$$

$$\sum_i\sum_j\sum_k \alpha^*_{i,j,k}\gamma_1(i,c_1) \leq \delta\gamma^*_1+(1-\delta)\gamma_1(M,c_1),$$

$$\sum_i\sum_j\sum_k \alpha^*_{i,j,k}\gamma_2(j,c_2) \leq \delta\gamma^*_2+(1-\delta)\gamma_2(M,c_2),$$

$$0 \leq \alpha^*_{i,j,k} \leq \infty, \qquad i \leq M, j \leq M, k \leq W$$

$$-\infty \leq \beta_1, \beta_2 \leq 0$$

$$1 \leq \delta \leq \infty$$

Despite linearizing the objective function, the optimization model is difficult to solve especially on account of the large number of non-linear equality constraints that exist because of the first four

constrains. The form that $P_i(n_i, n_j, \bar{\beta})$ takes becomes important for computational tractability. Recall that we express the workload experienced by nurses within a ward as a function of the nurse-patient ratio as

$$\gamma_i(n_i, c_i) = f(\frac{n_i}{c_i})$$

If we assume that the function $\gamma_i(\cdot) \forall i$ is linear, the marginal increase in workload experienced by nurses in any ward i on addition of a patient in equation 1 simplifies to

$$\Delta_i(n_i, c_i) = \frac{1}{c_i}$$

Our proposed routing policy now takes the form

$$P_i(n_1, n_2, \bar{\beta}) = \frac{(1/c_i)^{\beta_i}}{\sum_j (1/c_j)^{\beta_j}}$$

Where $P_i(n_1, n_2, \bar{\beta})$ is the probability that an incoming patient is assigned to ward i when there are n_1, n_2 patients currently in wards 1 and 2. Since the number of nurses in the ward is assumed to be a predetermined constant, we can rewrite $(1/c_i)^{\beta_i} = y_i$ following which our routing policy simplifies to

$$P_i(n_1, n_2, \bar{\beta}) = \frac{w_i}{\sum_j w_j} = y_i$$

We note that the probability of a patient being routed is now independent of the number of patients in a the wards. The optimization model now simplifies to

$$\underset{\bar{\alpha^*},\bar{\beta}}{\text{minimize}} \quad \frac{\sum_{i,j,k}[i+j+k]\alpha^*_{i,j,k}}{\lambda}$$

subject to $\alpha^*_{i,j,k}[j\mu_2(j,c_2)+i\mu_1(i,c_i)+\lambda] = [i+1]\mu_1(i+1,c_1)\alpha^*_{i+1,j,k}$

$$+[j+1]\mu_2(j+1,c_2)\alpha^*_{i,j+1,k}+\lambda y_1\alpha^*_{i-1,j,k}+\lambda y_2\alpha^*_{i,j-1,k}, \qquad i<M, j<M, k=0$$

$$\alpha^*_{i,j,k}[j\mu_2(j,c_2)+i\mu_1(i,c_1)+\lambda] =$$

$$+[j+1]\mu_2(j+1,c_2)\alpha^*_{i,j+1,k}+\lambda y_1\alpha^*_{i-1,j,k}+\lambda y_2\alpha^*_{i,j-1,k}, \qquad i=M, j<M, k=0$$

$$\alpha^*_{i,j,k}[j\mu_2(j,c_2)+i\mu_1(i,c_1)+\lambda] =$$

$$+[i+1]\mu_1(i+1,c_1)\alpha^*_{i+1,j,k}+\lambda y_1\alpha^*_{i-1,j,k}+\lambda y_2\alpha^*_{i,j-1,k}, \qquad i<M, j=M, k=0$$

$$\alpha^*_{i,j,k}[j\mu_2(j,c_2)+i\mu_1(i,c_1)+\lambda] = i\mu_1(i,c_1)\alpha^*_{i,j,k+1}$$

$$+j\mu_2(j,c_2)\alpha^*_{i,j,k+1}+\lambda[\alpha^*_{i-1,j,k}+\alpha^*_{i,j-1,k}], \qquad i=M, j=M, k=0$$

$$\alpha^*_{i,j,k}[j\mu_2(j,c_2)+i\mu_1(i,c_1)+\lambda] = i\mu_1(i,c_1)\alpha^*_{i,j,k+1}$$

$$+j\mu_2(j,c_2)\alpha^*_{i,j,k+1}+\lambda\alpha^*_{i,j,k-1}, \qquad i=M, j=M, 0<k<W$$

$$\sum_{(i,j,k)\neq(M,M,W)} \alpha^*_{i,j,k} = 1$$

$$\sum_i\sum_j\sum_k \alpha^*_{i,j,k}\gamma_1(i,c_1) \leq \delta\gamma_1^* + (1-\delta)\gamma_1(M,c_1),$$

$$\sum_i\sum_j\sum_k \alpha^*_{i,j,k}\gamma_2(j,c_2) \leq \delta\gamma_2^* + (1-\delta)\gamma_2(M,c_2),$$

$$0 \leq \alpha^*_{i,j,k} \leq \infty, \qquad i \leq M, j \leq M, k \leq W$$

$$0 \leq y_1, y_2 \leq 1$$

$$1 \leq \delta \leq \infty$$

The difficult constraints in the model are now simplified to be bi-linear equality constraints. While these are still not the easiest to handle, past literature has explored ways of relaxing them and coming up

with suitable approximations [Tho20; McC76]. One of these approximations is Mccormick relaxation [McC76] which is a type of convex relaxation used in bi-linear non linear programming problems. It involves transforming the non convex function into a convex function by relaxing the bi-linear variables. While this decreases the computational difficulty of solving the problem, it comes at a cost of introducing solutions that don't correspond to the original optimization problem. It becomes important, thus, to chose a convex relaxation that has the tightest bounds possible. A number of authors have worked on tightening the bounds when performing the Mccormick relaxation [Cas15; Naj17]. The general idea behind using this technique is as follows. Let us assume a non-linear programming problem that has the following bi-linear term xy, with $x^L \leq x \leq x^U$ and $y^L \leq y \leq y^U$, where x^U, x^L, y^U, y^L are the upper and lower bounds for decisions variables x and y respectively. The relaxation procedure involves introducing a new term w in place of xy and introducing the following set of constraints

$$w \geq x^L y + xy^L - x^L y^L$$

$$w \geq x^U y + xy^U - x^U y^U$$

$$w \geq x^U y + xy^L - x^U y^L$$

$$w \geq x^L y + xy^U - x^L y^U$$

In this way, introducing convexity into the problem allows us to obtain a global optimum, which is obtained to be a lower bound to the original problem. Additionally, using spatial branch and bound methods [Tuy98] allows for the creation of new lower bounds that are closer and closer to the true solution to our original problem. To reformulate the optimization problem using McCormick envelops, the term $y_l \alpha^*_{i,j,k}$ is replaced with $x^l_{i,j,k}$. Additionally, we add the 4 constraints representing the McCormick envelopes for each $x^l_{i,j,k}$. To formulate the constraints representing these envelopes we require knowledge of the upper and lower bounds for y_l and $\alpha^*_{i,j,k}$. Recall that the range for y_l and $\alpha^*_{i,j,k}$ is $[0,1]$ and $[0,\infty]$ respectively. Using a large value M^∞ to represent infinity, the constraints representing the McCormick envelopes for variables y_l, $\alpha_{i,j,k}$ and $x^l_{i,j,k}$ can be written as

$$x^l_{i,j,k} \geq 0$$

$$x^{l}_{i,j,k} \geq \alpha_{i,j,k} + M^{\infty} y_l - M^{\infty}$$

$$x^{l}_{i,j,k} \geq M^{\infty} y_l$$

$$x^{l}_{i,j,k} \geq \alpha_{i,j,k}$$

A potential extension to the work done in Chapter 3 would be to study the performance of the relaxed optimization problem as we increase the problem size by adding more wards and allowing for higher capacity within each ward. The complete reformulated optimization model can now be written as follows

$$\underset{\alpha_{i,j,k},y_1,y_2}{\text{minimize}} \quad \frac{\sum_{i,j,k}[i+j+k]\alpha_{i,j,k}}{\lambda[1-\alpha_{M,M,W}]}$$

subject to

$$\alpha_{i,j,k}[j\mu_2(j,c_2)+i\mu_1(i,c_i)+\lambda] = [i+1]\mu_1(i+1,c_1)\alpha_{i+1,j,k}$$
$$+[j+1]\mu_2(j+1,c_2)\alpha_{i,j+1,k}+\lambda x^1_{i-1,j,k}+\lambda x^2_{i,j-1,k}, \qquad i<M, j<M, k=0$$

$$\alpha_{i,j,k}[j\mu_2(j,c_2)+i\mu_1(i,c_1)+\lambda] =$$
$$+[j+1]\mu_2(j+1,c_2)\alpha_{i,j+1,k}+\lambda x^1_{i-1,j,k}+\lambda x^2_{i,j-1,k}, \qquad i=M, j<M, k=0$$

$$\alpha_{i,j,k}[j\mu_2(j,c_2)+i\mu_1(i,c_1)+\lambda] =$$
$$+[i+1]\mu_1(i+1,c_1)\alpha_{i+1,j,k}+\lambda x^1_{i-1,j,k}+\lambda x^2_{i,j-1,k}, \qquad i<M, j=M, k=0$$

$$\alpha_{i,j,k}[j\mu_2(j,c_2)+i\mu_1(i,c_1)+\lambda] = i\mu_1(i,c_1)\alpha_{i,j,k+1}$$
$$+j\mu_2(j,c_2)\alpha_{i,j,k+1}+\lambda[\alpha_{i-1,j,k}+\alpha_{i,j-1,k}], \qquad i=M, j=M, k=0$$

$$\alpha_{i,j,k}[j\mu_2(j,c_2)+i\mu_1(i,c_1)+\lambda] = i\mu_1(i,c_1)\alpha_{i,j,k+1}$$
$$+j\mu_2(j,c_2)\alpha_{i,j,k+1}+\lambda\alpha_{i,j,k-1}, \qquad i=M, j=M, 0<k<W$$

$$\sum_{i,j,k}\alpha_{i,j,k}=1$$

$$\sum_i\sum_j\sum_k\alpha^*_{i,j,k}\gamma_1(i,c_1) \le \delta\gamma^*_1+(1-\delta)\gamma_1(M,c_1),$$

$$\sum_i\sum_j\sum_k\alpha^*_{i,j,k}\gamma_2(j,c_2) \le \delta\gamma^*_2+(1-\delta)\gamma_2(M,c_2),$$

$$x^l_{i,j,k} \ge \alpha_{i,j,k}+y_l-1, \qquad i\le M, j\le M, k=0, l=\{1,2\}$$

$$x^l_{i,j,k} \ge \alpha_{i,j,k}, \qquad i\le M, j\le M, , k=0, l=\{1,2\}$$

$$x^l_{i,j,k} \ge y_l, \qquad i\le M, j\le M, , k=0, l=\{1,2\}$$

$$0\le\alpha_{i,j,k}, y_1, y_2\le 1, \qquad i\le M, j\le M, k\le W$$