Strategic and tactical decision-making for inpatient admission and hospital bed allocation: an application to neurology wards

Saied Samiedaluie

Doctor of Philosophy

Desautels Faculty of Management

McGill University

Montreal, Quebec

October 2014

A thesis submitted to McGill University
in partial fulfillment of the requirements of the degree of
Doctor of Philosophy

DEDICATION

To my precious parents, Mahin and Mohammadreza, for their unwavering support and encouragement.

ACKNOWLEDGMENTS

First and foremost, I would like to express my special appreciation and sincere gratitude to my advisor, Dr. Vedat Verter. With his immense knowledge and unconditional support, Dr. Verter made my Ph.D. study a very wonderful and rewarding journey. His guidance and advice, as well as his wonderful personality, created a pleasant experience, every moment of which was tremendously enjoyable. I am also extremely indebted to my former co-advisor, Dr. Dan Zhang, for his enthusiasm, his trust in my abilities, and his efforts in keeping me motivated. His generous support and invaluable advice never stopped even after he was not officially my advisor. Meanwhile, I am very grateful to the other members of my Ph.D. committee, Dr. Beste Kucukyazici and Dr. Michele Breton, for their brilliant comments and suggestions. I would like to thank all the Operations Management professors, the staffs at the Ph.D. program office, and my Ph.D. fellows at the Desautels Faculty of management, who were always a great source of friendship and help. I especially thank my parents, sisters, and brothers. My hard-working parents have sacrificed their lives for their children and provided unconditional love and care. I love them so much, and I believe I would not have made it this far without my family's support. Finally, a big thank you to my dear friends in Montreal, who were always there for me. We had marvelous memories together, which I will cherish for the rest of my life.

CONTRIBUTIONS OF AUTHORS

Two manuscripts have been written based on this thesis:

1. *Samiedaluie, Saied, Beste Kucukyazici, Vedat Verter, Dan Zhang. 2013. "Managing Patient Admissions in a Neurology Ward". Under second revision for resubmission to Operations Research.*

The student (first co-author) developed the modeling framework and solution methodology, reviewed the literature, and designed the computational analysis with the guidance provided from other co-authors. The second co-author collected and analyzed the data and helped in defining the research problem. The third co-author contributed in modeling the problem, designing the computational analysis, and interpreting the results as well as editing the manuscript. The fourth co-author was actively involved in formulation of problem and devising the solution methodology. During the whole process the first author was the main person responsible for conducting the research, writing the manuscript, and developing the computer programs and simulation models using comments and feedback from co-authors.

2. *Samiedaluie, Saied, Vedat Verter. 2014. "Does Specialization of Healthcare Services Improve Operational Efficiency?" Working Paper.*

The student (first co-author) and the second co-author collectively defined the research question and developed the study framework. The first co-author reviewed the literature, formulated the problem, developed solution algorithms and simulation models, and designed the experiments with the guidance of second co-author. The

second co-author provided editorial help and recommendations for problem formulation and computational analysis.

ABSTRACT

The healthcare sector has received constant appeals from different stakeholders over the past decades to increase operational efficiency and enhance quality of care for patients. Hospital managers and health authorities confront serious challenges in identifying areas for improvement and designing plans to boost healthcare delivery processes, all while maintaining the operational costs aligned with their planned budget. The difficulty of this task is amplified by budget cuts and insufficient resources in the healthcare system. Operations Research models can be used to assist healthcare managers in making informed and evidence-based decisions. This thesis aims at developing patient admission and bed allocation policies in acute care wards, where acquiring extra resources is extremely expensive for hospitals and a delay in treatment is highly undesirable from a patient health perspective.

The problem of patient admission and inpatient bed allocation in acute care wards recognizing multiple patient types with different medical characteristics is considered in this thesis. Recent studies have shown that in the event of an acute episode patients are more effectively treated in specialized inpatient settings. The benefits of such specialized care, however, might be offset by long wait times at the emergency department due to bed unavailability in the ward. This research is inspired by the managerial challenges at the neurology ward of the Montreal Neurological Hospital, where the optimal care pathway for patients with neurological diseases is particularly time-sensitive. Failure in matching the hospital's service capacity and patient demand for certain levels of care can be problematic. Moreover, day-to-day

fluctuations in demand affect the efficient utilization of hospital capacity. The key issue for matching the demand and service capacity and improving the performance of the hospital is intelligently designed capacity-related policies; both at the strategic and tactical levels.

At the tactical level, the admission process of patients to a neurology ward is modeled using an average cost dynamic programming framework. By solving the dynamic program model, we are essentially looking for the dynamic admission policy that provides the best care for all patients in light of limited bed availability. In terms of solution methodology, an integrated approach that combines queuing models and approximate dynamic programming is presented. Furthermore, the performance of the proposed approach is compared with the performance of other heuristic policies that can be suggested for such types of problems. It is shown that the dynamic admission policy that can adjust allocations of the beds based on the state of the ward performs better compared with other static policies. In particular, the dynamic admission policy reduces the average ED boarding time that patients experience before they are transferred to the ward.

At the strategic level, the problem of multi-site resource allocation and system configuration in response to the pending merger of two existing sites, i.e., the stroke wards at Montreal Neurological Hospital and Montreal General Hospital, is studied. Designing an appropriate admission policy for patients at the hospital level along with an optimal bed allocation policy between the two sites are the major concerns of hospital managers in this process. Two possible settings for admission of patients to the hospitals are examined to determine which setting would be preferred

in terms of minimizing the patient admission refusal rate. Meanwhile, the multi-site bed allocation problem is formulated so that resources are optimally distributed in accordance with the patient flow at each site. It is found that the decision of system configuration for a multi-hospital network requires careful consideration of patient mix in the arrivals, relative length of stay of patients, and the distribution of patient load between hospitals.

RÉSUMÉ

Le secteur de la santé a reçu des appels constants des différentes parties prenantes au cours des dernières décennies pour améliorer l'efficacité opérationnelle et la qualité des soins pour les patients. Les gestionnaires d'hôpitaux et autorités de santé sont confrontés à des problèmes graves et doivent identifier les domaines d'amélioration et concevoir des plans d'amélioration des processus de prestation de soins de santé, tout en maintenant les coûts opérationnels dans les limites de leur budget. La difficulté de cette tâche est amplifie par les compressions budgétaires et l'insuffisance des ressources dans le système actuel de soins de santé. Modèles de la recherche opérationnelle peuvent être utilisés pour aider les gestionnaires de soins de santé à prendre des décisions éclairées et fondées sur des données probantes. Cette thèse vise à développer des politiques d'admission et d'allocation de lits des patients dans les services de soins de courte durée, où l'acquisition de ressources supplémentaires est extrêmement coûteuse pour les hôpitaux, et où retard dans le traitement est hautement indésirable du point de vue de la santé des patients.

Cette thése s'intèresse au problème de l'admission de patients et de l'attribution des lits en milieu hospitalier dans les services de soins de courte durée, tenant compte de la multiplicité des types de patients et de leurs caractéristiques médicales spécifiques. Des études récentes ont montré que les patients présentant un épisode aigu sont plus efficacement traités en milieu hospitalier spécialisé. Les avantages d'une telle spécialisation des soins peuvent cependant être atténués par des longs délais d'attente

à l'urgence en raison de l'indisponibilité de lits dans la salle. Cette recherche est inspire par les défis de gestion au service de neurologie de l'Hôpital neurologique de Montréal, où le parcours de soins optimal pour les patients atteints de maladies neurologiques est particulièrement. Une insuffisance de la capacité de service de l'hôpital pour répondre à la demande de patients pour certains niveaux de soins peut être problématique. En outre, les fluctuations journaliéres de la demande influent sur l'utilisation efficace de la capacité de l'hôpital. La solution à l'adaptation de la capacité à la demande et à l'amélioration de la performance de l'hôpital réside dans la mise au point de politiques intelligentes en matière de gestion de capacité, tant au niveau stratégique que tactique.

Au niveau tactique, le processus d'admission des patients à un service de neurologie est modélisé par un programme dynamique de minimisation du coût moyen. En résolvant le modèle programme dynamique, nous cherchons essentiellement la politique d'admission dynamique qui offre les meilleurs soins pour tous les patients à la lumière de la disponibilité limitée des lits. En termes de méthodologie de solution, nous utilisons une approche intégrée qui combine des modèles de files d'attente et la programmation dynamique approximative. En outre, la performance de l'approche proposée est comparée à la performance d'autres politiques heuristiques suggérées pour ce type de problèmes. Nous montrons que la politique d'admission dynamique ajustant l'allocation des lits sur la base de l'état du système présente de meilleures performances par rapport aux politiques statiques. En particulier, la politique d'admission dynamique réduit le temps moyen d'attente des patients à l'urgence avant qu'ils ne soient transférés en neurologie.

Au niveau stratégique, nous examinons le problème de l'allocation des ressources sur des sites multiples et la configuration du système en réponse à la fusion en cours des deux sites existants, à savoir l'Hôpital neurologique de Montréal et l'Hôpital général de Montréal. Les préoccupations principales des gestionnaires d'hôpitaux dans ce processus sont la conception d'une politique d'admission approprie pour les patients au niveau de l'hôpital et d'une politique optimale de l'attribution des lits entre les deux sites. Deux configurations possibles pour l'admission des patients dans les hôpitaux sont examines pour déterminer les meilleurs paramètres en termes de minimisation du nombre de refus d'admission de patients. Par ailleurs, le problème de l'attribution des lits multi-sites est formulé de telle sorte que les ressources soient distribuées de façon optimale en fonction du flux des patients à chaque site. Nous constatons que la décision de configuration du système pour un réseau multi-hôpital nécessite un examen attentif de la composition des arrivées de patients, de la longueur relative de séjour des patients, et de la répartition du nombre de patients entre les hôpitaux.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

1.1. Inpatient Admission and Bed Allocation in Hospitals

Resources are the vital components of any production or service system. In hospitals, the resources that managers use for making optimal capacity decisions are inpatient beds, staff (including nurses and physicians), departments (including Emergency Departments, Operating Rooms, and Laboratories), medical equipment (like MRI and CT scan), etc. This thesis focuses on managing the inpatient beds, which extensively impact the operation of a hospital. As an emphasizing fact, the number of properly staffed inpatient beds is an important factor in determining hospital capacity (Green, 2006).

The number of appropriately staffed beds determines the requirements of other resources in hospitals as well. For example, using nurse-to-bed ratios defined for providing care to a specific cohort of patients in a ward, the nurse staffing levels are decided. Hospital managers can adjust the service capacity of their hospitals and the associated resource requirements by opening or closing the inpatient beds. Throughout this thesis, whenever the "inpatient bed" is mentioned it implicitly refers to appropriately staffed inpatient beds.

Hospital managers face various challenges in making the right decisions related to management of inpatient beds and admission of patients in healthcare systems. There is continuous pressure from different stakeholders to optimize the healthcare

delivery process and enhance the quality of care. Furthermore, constant budget cuts in the healthcare sector force hospital managers to decrease the operational cost and keep it aligned with planned financial resources. Therefore, it is imperative that the decisions related to inpatient beds, as crucial resources in the hospital, and admission of patients, as the primary consumers of resources, are accompanied with comprehensive analysis of their consequences and impacts on the performance of the hospitals.

Capacity management decisions regarding inpatient beds in healthcare systems are classified into two broad categories: capacity acquisition and capacity allocations decisions (Smith-Daniels et al., 1988). Capacity acquisition decisions are concerned with determining the right number of beds needed to meet some predefined goals while capacity allocation decisions look for efficient ways of using available resources. Choosing the right number of beds in a hospital is not an easy task to accomplish since providing sufficient and timely care to the patients and avoiding spending more than necessary resources should be taken into consideration simultaneously. Even though this sort of decisions is more likely to be made within a hospital, one should bear in mind that it might be seen at the regional level, mostly in publicly-funded healthcare systems. Capacity acquisition decisions are considered to be *strategic* decisions, since it is hard to revise them frequently once they are made and implemented.

On the other hand, capacity allocation decisions usually answer the following question: How should we distribute the available resources across different departments in a hospital or across different types of patients to achieve our targets? When making allocation decisions, it is assumed that the available resources are fixed. For

instance, let us assume there is a limited number of beds in a ward and multiple types of patients are waiting to be admitted to the beds. Then the task of assigning beds in a way to minimize the waiting times while maintaining equity among different types of patients is an example of allocation decisions. Taking into account stochastic arrivals and different lengths of stay for each type of patients adds on the complexity of the problem. The capacity allocation decisions are considered to be at the *tactical* level and can be modified in short term.

One of the early surveys which reviews, classifies and analyzes the literature on capacity management in healthcare is Smith-Daniels et al. (1988). This survey deals with all types of resources in a hospital, but their classification provides an appropriate overview on capacity decisions in hospitals and is therefore presented in Table 1–1. The topic of this thesis relates to determining the size of inpatient units and inpatient admissions (highlighted items in Table 1–1).

Table 1–1: Classification of capacity management themes by Smith-Daniels et al. (1988)

	Facility (Physical) Resources	Workforce Resources
Acquisition Decisions	Facility location and capacity size	Hospital Staffing
	Size of inpatient care units	Ambulatory care staffing
	Size of ambulatory care units	
Allocation decisions	**Inpatient admissions scheduling**	Assign workers to shifts
	Surgical facility scheduling	Assign workers to units
	Ambulatory care scheduling	Assign workers to tasks

1.2. Strategic and Tactical Capacity Decision-Making

Failure in matching the hospital's service capacity and the patients' demand for certain levels of care can be problematic. Moreover, day-to-day fluctuations

in demand affect the efficient utilization of hospital capacity. The key issue for matching the demand and service capacity and improving the performance of the hospital is intelligently-designed capacity-related policies; both at the *strategic* and *tactical* levels.

At the *tactical* level, the patient admission and bed allocation problem in an acute care ward recognizing multiple patient types with different medical characteristics is studied in this thesis. The manager of the acute care ward decides on the admission of patients from the Emergency Department (ED) to the ward on a daily basis – or hourly basis, depending on the congestion of the system. This requires designing policies according to which patients are prioritized or the available beds are allocated to different patient types. These policies are either static, which means the rules do not change over time, or dynamic, which implies the recommended decisions depend on time or the state of the system.

At the *strategic* level, the multi-site bed allocation problem and multi-hospital network configuration design are considered. Mergers of hospitals and developing healthcare networks to collectively provide care to patients have become very popular recently. In this process, the configuration of network and devising the patient admission policies at the level of hospitals are the main questions to be answered. Further, the optimal inpatient bed allocation to hospitals in accordance with the flow of patients needs be addressed. The resource allocations are at the level of hospitals and revising such decisions are quite costly in these situations.

1.3. Bed Capacity Issues in Montreal Neurological Hospital

The application of strategic and tactical decision-making problems pertinent to inpatient beds and patient admissions are studied with the collaboration of the Montreal Neurological Hospital (MNH). The managerial issues at the neurology ward of this hospital are addressed in this thesis. The MNH is an academic medical center that provides care to patients suffering from neurological diseases. Neurological diseases, including ALS, Alzheimer, Multiple Sclerosis, Parkinson's, spinal cord injury and stroke, represent leading causes of death and disability in the Canadian and US populations (World Health Organization, 2006).

Many neurological conditions are chronic, worsen over time and produce a range of functional limitations posing daily challenges to patients and their caregivers. For example, Heart and Stroke Foundation (`www.heartandstroke.com`) identifies stroke as the third leading cause of death in Canada with about 14,000 fatalities each year; and reports that about 300,000 Canadians are living with the effects of stroke. The Global Burden of Disease study conducted in 2002 by the World Health Organization also determined that neurological conditions accounted for 38.3% of the disability-adjusted life years (DALYs) worldwide (Lopez et al., 2006), while the percentages observed in developed countries are much higher than the global average.

Recent studies have shown that such critically ill patients are more effectively treated in specialized inpatient settings offering properly organized care (Chalfin et al., 2007). The benefits of such specialized care, however, might be offset by long wait times at the ED due to bed unavailability at the inpatient ward. This research is inspired by the managerial challenges at the neurology ward at MNH, where the

optimal care pathway for patients with neurological diseases is particularly time-sensitive (Castillo, 1999).

Currently, a fixed number of beds in the neurology ward of MNH have been dedicated to each type of patients and some extra beds are assigned to be used by all types. We will look for a dynamic admission policy that provides the best care for all patients in light of limited bed availability. The dynamic admission policy that can adjust allocations of the beds based on the state of the ward is believed to perform better compared with other static policies. Designing the patient admission policies is identified as a *tactical* decision-making problem.

As a more *strategic* managerial challenge, two existing sites, i.e., the stroke wards of MNH and Montreal General Hospital (MGH) are pending to be merged and administered centrally. The proposed structural change requires each hospital to provide one of the two levels of care (secondary and tertiary) needed by stroke patients. This implies that each site will be dedicated to serve only a specific type of patients. As an alternative structure, the two hospitals can provide all levels of care and accommodate all types of patients.

Designing an appropriate configuration of system along with an optimal bed allocation policy between the two sites are the major concerns of hospital managers in this process. Two possible settings for patients' admissions to the system are examined to determine which setting would be preferred in terms of minimizing the patient admission refusal rate. Meanwhile, the multi-site bed allocation problem is formulated so that the resources are optimally distributed in accordance with the patients flow in each site.

1.4. Contributions of This Research

This study aims at exploring the possibility of using dynamic admission policies in acute care wards and examining their impact on improving the operational performance of wards and quality of treatment outcome from patient health perspective. It illustrates how a dynamic admission policy is derived with the help of mathematical models taking into account the operational settings of every acute care ward and medical characteristics of inpatients. Given the complexity associated with the outputs of mathematical analysis, the process of converting the results to practical and easy-to-implement recommendations is also provided.

Another contribution of this research in the context of healthcare operations management is evaluating the benefits of narrowing down the scope of care in a multi-hospital network compared with keeping the scope of care broad. The hospitals that are specialized in providing a specific type of care benefit from the economies of focus. However, such benefits might diminish by losing the economies of scale. This research sheds additional light on the problem of system configuration in networks where specialization and diversification of healthcare services are the possible scenarios. It provides information on the characteristics of multi-hospital networks that are crucial in deciding which scenario is the recommended alternative.

From a methodological perspective, an application of approximate dynamic programming in healthcare is presented in this research. The approximating approach provides an idea of combining queueing theory and an LP-based approximate dynamic program. It creates a framework for decomposing the original dynamic program, which is not solvable in a reasonable amount of time for large-scale instances

of the problem, into smaller sub-programs that are efficiently solved with standard techniques such as the value iteration algorithm. For the multi-hospital network design study, the multi-site bed allocation problem is formulated and a heuristic algorithm is developed to find the best allocation of beds between the hospitals in each configuration scenario.

1.5. Organization of the Thesis

The literature related to the problems considered in this thesis is reviewed in Chapter 2. In Chapter 3 the problem of patient admission for acute care wards with an application to a neurology ward is studied. Chapter 4 examines the idea of specialization of healthcare services in the framework of a multi-hospital network configuration problem. Chapter 5 concludes the thesis with the findings and contributions of this research to the knowledge of managing inpatient beds and patient admission policy design.

CHAPTER 2
Literature Review

In this section, we review the operations research and management science literature related to the bed allocation and patient admission problem in hospitals. These problems have received attention in the academic literature for more than four decades. In order to present a structured review, we have identified four categories under which the related studies are summarized.

2.1. Regional Hospital Bed Planning

Ruth (1981) studies resource allocation problem between the hospitals that are located in a region. The demand for hospital care is classified by geographical area and level of care. Furthermore, the level of care at each hospital and its conformance are determined based on the number of beds allocated to that hospital. To find the new optimal configuration of resources in the region, a mixed-integer program is developed so that the total cost of system modification is minimized while some constraints are satisfied. These constraints are; accessibility constraints, which guarantee that the population in each region is served; acceptability constraints, which ensure that the increased level of care in a hospital will be used by people; and conformance constraints that take care of feasibility of making a hospital conforming.

In a more recent study related to regional hospital bed planning, Santibáñez et al. (2009) develops a multi-period mathematical model that helps health authorities allocate hospital resources in a region and reconfigure the hospital network to

accommodate projected demand in the future. The model provides solutions for locating clinical services across the hospitals and assigning the required resources while considering all clinical and operational restrictions such as minimum annual patient visits per hospital, ratio of doctors-to-population in a region, standard size of clinical units, etc. The main objective of the model is to maximize the access to care for all populations in each region and minimize the efforts of relocating services over all periods in the planning horizon. The model has been applied to a hospital network in British Columbia including 12 hospitals to figure out the structure of healthcare delivery for the next 15 years.

Güneş and Yaman (2010) studies the problem of merging two hospital networks that requires reallocation of resources and redefining the service portfolios of hospitals. By deciding which services should be provided in each hospital, the required resources including beds, doctors, and nurses are also determined. The objective of the problem is to minimize the cost of transferring resources between hospitals along with the cost of patients' transportation that is incurred as a result of restructuring.

Abdelaziz and Masmoudi (2011) develops a multi-objective stochastic program to find the number of beds in a network of hospitals that are needed to provide three levels of care to patients; primary, secondary and tertiary care. The demand for primary care should be satisfied by the hospital where the patient is observed. The demand for secondary and tertiary care can be satisfied at the regional and national level, respectively. The level of care in each hospital is determined based on the specialties the hospital offers. Identifying the specialties in each hospital is part of the decision-making process in this study. Based on the number of beds at each

department and specialties in each hospital, the number of physicians and nurses is determined. The first objective of the stochastic model is to minimize the cost of creating new beds and minimizing the number of nurses and physicians are the second and third objectives.

2.2. Bed Allocations within a Hospital

One of the early simulation studies of the bed allocation problem is Vassilacopoulos (1985). The simulation model developed in this paper is used to determine the number of beds in different inpatient departments with the objective of balancing some operational efficiency measures. Admission of emergency patients with no delay, keeping the occupancy level above and the waiting lists below specified thresholds are the important efficiency measurements in this study. Using the simulation model, several bed allocation policies as well as changing the mix of schedule and emergency patients has been evaluated.

Based on hourly census data, Lapierre et al. (1999) develops a time-series model that assists with the allocation of beds between different medical units within a hospital. Using this model, hospital administrators can decide how many beds should be assigned to each unit to have the same number of bed shortage occurrences across the units. Kusters and Groot (1996) also uses statistical methods to support the decision-making process related to bed allocation. In a different study, Cochran and Roche (2007) examines using different inpatient data sets in order to calculate the bed demands across all levels of care in a hospital. This study concludes that the financial and billing data sets are more appropriate for using in estimating the hospital's demands for beds. For allocating beds to multiple level of care, the authors

use queueing models to calculate the utilization rate and then by adjusting the number of beds, they try to meet the utilization targets of each unit.

Green (2002) discusses the issue of determining the hospital capacities based on target levels of occupancy. As an alternative criterion, the author focuses on the probability of having a bed available for a new patient. The issue of bed unavailability and its medical consequences on patients as well as operational impact on the other units of hospital, especially ED, is a great concern in hospital units where timeliness of care is a priority. By using a very simple queueing model, the probability that a patient has to wait for an inpatient bed is calculated in this paper. Moreover, the relationship of occupancy levels and probability of bed unavailability with number of beds have been explored and the provided insight can be used to decide the capacity needs. Another study that emphasizes on the importance of using appropriate measures of delay in determining bed requirements is Green and Nguyen (2001). Using queueing theory, the authors examine the effect of reducing beds, increasing the demand, consolidating the clinical units, and reducing LOS. One of the key findings of this study is that reducing the mean LOS has greater potential to reduce the bed needs than its variability. Also, consolidation of clinical units might increase the patients delay depending on the relative arrival rates and LOS of different patient types.

Harper and Shahani (2002) develops a flexible simulation model that captures the patients flow in a hospital from arrival to discharge. The patients are admitted to the units based on their priority listings and if a bed is not available in the first choice of the patients, another bed in other units in the list will be chosen. The admission of

elective patients can be deferred and those patients might be asked to come back in a few days if all beds are full. The model incorporates the necessary details related to variability in demand and LOS, which are very important for effective bed planning in a hospital. The model is useful in estimating the bed occupancies and admission refusal rates as a consequence of any changes that the managers plan to do in terms of restructuring and reallocating the beds between different units.

Hospitals can periodically change allocation of beds between different units to respond to anticipated demand. Kao and Tung (1981) proposes an approach to solve such problem with the objective of minimizing the overflow of patients between units. The demand is forecasted through an $M/G/\infty$ queue and the minimum number of beds needed to accommodate a prescribed amount of patients load over a year is determined. The rest of the beds in the hospital are distributed over different services to minimize the expected average overflows during the months of the planning horizon. In a later study, Akcali et al. (2006) considers the same problem over a finite horizon. The number of allocated beds can change over time, but there is some cost associated with shifts in capacity. In addition to the constraint on budget for changing capacity in each period, the waiting time of patients is also bounded by a maximum allowed limit. The problem has been formulated as a non-linear mixed integer program, which under some realistic assumptions is solved in a reasonable amount of time.

In an integrated model of queueing theory and goal programming, Li et al. (2009) formulates the bed allocation decisions across different departments of a hospital, which takes into account the objectives of each department related to patient service

level and profits. Bretthauer et al. (2011) develops a model to determine the optimal number of beds in different units of a hospital to minimize the blocking probabilities with respect to a budget constraint. The flow of a patient is blocked if there is no bed available in the next unit of that patient's care path. The authors present a heuristic algorithm that estimates the blocking probability in a queueing network and demonstrates its superiority over existing heuristics.

Garg et al. (2010) develops a Markov chain to model patients flow through hospital and community care phases. The model can be used to forecast resource requirements as an input for patient scheduling or resource allocation planning. Osorio and Bierlaire (2009) models the congestion in a network of hospital units that results in blocking the patients from moving through phases of clinical care. The presented analytical framework considers finite capacity for queues and takes into account the between-queue correlations to describe the sources of congestion and its impact on the network.

2.3. Inpatient Bed Management in a Hospital Ward

Queueing models are effective tools in measuring the performance of a single hospital ward and evaluating the impacts of capacity decisions on the system. Two papers of Gorunescu et al. (2002a) and Gorunescu et al. (2002b) use queueing models to study resource allocations in a geriatric department. In Gorunescu et al. (2002a), the authors consider an $M/PH/c$ queue that assumes phase-type distributions for the LOS of patients in the ward. They also assume that the patient will be lost if all the beds are occupied. Using this queue, they calculate the fraction of patient arrivals that are turned away because of bed unavailability. Therefore, with a maximum

acceptable delay probability, they can specify the minimum number of beds that are needed to achieve that target. In order to balance the service level for patients and utilization of resources, they consider the cost of having an empty bed to meet the demand versus the cost of turning away patients. In a follow-up work, Gorunescu et al. (2002b) introduce extra unstaffed beds to be used in the case of demand surges. They use an $M/PH/c/N$ queue to calculate the effect of changes in the arrival rate, LOS and bed allocations on rejection probabilities, utilization and costs of the system.

Utley et al. (2003) studies the problem of having an intermediate care unit in a hospital for admitting the patients who do not need acute care. Based on a mathematical model that captures the patients flow between acute and non-acute units, the author estimate the percentage of days that shortage occurs, either in acute or non-acute units. However, a major drawback of this model is that when the demand for beds exceeds capacity, no assumptions are made regarding its effect on the arrival or flow of patients.

Ayvaz and Huh (2010) considers two types of patients arriving at a hospital: The patients that wait until they are served and the other type of patients who leave the system if they are not immediately accommodated. It is assumed that patients use only one unit of capacity and the occupied capacity will be released at the end of the day regardless of the time the patient has been admitted. Therefore, at the beginning of every day, all the capacity becomes available. To find the optimal number of admissions per day, a discounted total cost dynamic program has been developed and to solve the model, the authors proposed a heuristic policy that protects some

portion of capacity for those patients that leave the system if they are not served upon arrival.

Bekker and de Bruin (2010) examine the effect of timed-dependent patient arrivals on the number of required beds and the fraction of patients that are rejected. The impact of time varying arrival pattern of patients is influenced by the average length of stay (ALOS). If the ALOS is rather large respect to the cycle of arrival change, then the variation in the arrivals is averaged out. For example, in clinical wards with ALOS of two or three days, the daily variation does not affect the average number of beds. However, this might have effect on the emergency department functioning. The other key finding is that if the variability in LOS increases, it can stabilize the variation in the patient rejection rate in a loss queueing network.

2.4. Patient Admission Scheduling

Among the first studies, Kolesar (1970) develops a Markovian model that incorporates the scheduling of outpatients as well as the admission of inpatients that need immediate hospitalization. Queueing models are used to calculate the state probabilities of the system associated with different admission policies. A linear program based on these state probabilities is developed to determine the number of admissions for the next day. The objective of this linear program could be either maximizing the bed occupancy level while keeping the probability of refusing admission below a threshold or minimizing the admission refusal rate while keeping the occupancy level above a threshold. Esogbue and Singh (1976) considers the admission problem for two types of patients with a similar objective: maximization of occupancy and minimization of unsatisfied requests. Based on a cut-off priority policy, they develop

a birth and death process and solve the model for the optimal value of the cut-off priority.

Thompson et al. (2009) considers admission process of patients from the ED to different floors in a hospital. For each patient category, there is an ideal floor to be admitted to. However, there are some other floors that can be considered as an alternate admission destination. Admitting the patient to the ideal floor has lower cost (or larger reward) than admission to the alternate floor. The authors also assume that patients can be moved from one floor to another during their stay in the hospital, but there is some cost associated with that. They formulate the problem as a finite-horizon discrete-time Markov decision process (MDP) to find the best admission policies. Based on the optimization model, a decision support system has been developed and implemented in a hospital that showed significant increase in the revenue and drastic decrease in the average waiting time of the patients before admission. Modeling the same problem but from a different perspective, Mandelbaum et al. (2012a) models the problem of admitting patients from the ED to internal wards – named as patient routing – to find a fair and balanced workload for all units as well as improving the operational efficiencies. The applicability of different routing algorithms in the hospital in the light of data availability has also been discussed.

Demeester et al. (2010) develops a Tabu search to assign patients to different departments of a hospital according to their medical needs and personal preferences. Each department in a hospital is serving one major specialty, but can accept patients from a few other specialties as well. The objective of the model is to minimize the

number of patients that are admitted to departments other than the most suitable department.

Helm et al. (2011) examines the idea of having an expedited care queue in addition to emergency and scheduled patient queues in a hospital. The expedited care queue is designed to serve those scheduled patients who can wait for a few days to receive the care they need but choose the ED as gateway to get admitted to hospital. To find the optimal admission policy that balances the opportunity cost of unused resources with the cost of canceling scheduled patients and blocking emergency patients, an MDP has been developed. The insights obtained from the structural properties of the optimal policy in special cases provide intuitions for designing admission policies that are proved to work well using a simulation study.

CHAPTER 3
Managing Patients Admission in a Neurology Ward

3.1. Introduction

This chapter focuses on developing policies for admitting patients from ED to a neurology ward. Admission policies define the rules for allocating inpatient beds to multiple types of patients as well as transferring them to another hospital. In the event of an acute episode, neurology patients are admitted to hospital through the ED. Diagnosis of such conditions in the ED requires extensive physical examinations, often aided by a specialty team (such as a stroke team), brain imaging (CT or MRI) and other diagnostic tests. Neurology patients are more effectively treated in a specialized inpatient setting, i.e. the neurology ward.

The features of the neurology ward include; the care given by a specialized nursing team, the use of exclusively equipped beds, the availability of occupational, speech and physical therapies (Stroke Unit Trialists' Collaboration, 2007). As a result some patients' DALYs can be improved significantly through enhancing functional abilities. Thus, the accessibility to such specialized care setting is particularly time-sensitive for patients with acute conditions (Castillo, 1999). Indeed, Kucukyazici et al. (2010) observed that the potential benefits of the specialized care might be offset by long delays in the ED prior to admission to the neurology ward.

Recognizing the long-term negative effects of extended ED boarding on the health outcomes, the neurology patients may be transferred to another hospital,

which can offer such a specialized care for neurological diseases, following the diagnosis of the condition in the ED. This is a decision neurology ward managers strive to avoid since the patient faces additional waiting time at the transfer destination.

Many neurology wards face the problem of insufficient capacity to meet demand for inpatient beds, especially during demand surges. The problem is pronounced since admitting these patients to other wards is not an option, i.e., off-unit servicing is not feasible for these patients. Note that the capacity for patient care is determined not only by the number of beds in the neurology ward but also by the team of specialized nurses, physicians, and allied health professionals. The patient-to-specialized nurse and patient-to-neurologist ratios are key performance measures of quality of care. Moreover, the beds in these wards are specially equipped neurology beds and substitution of these beds by admitting those patients to other wards often has a negative impact on health outcomes.

A static patient admission policy is used in many neurology wards by assigning a fixed number of beds to each type of disease. Sometimes, a certain number of beds are used as flexible beds and shared among different types of patients. For example, at the MNH, there are sixteen beds in the neurology ward, where six beds are dedicated to stroke patients, six beds are dedicated to non-stroke neurology patients, and four of them are used as flexible beds to admit either stroke or non-stroke patients. It is important that neurology wards do not provide off-unit service to other wards in the hospital and hence, the neurology patients boarding at the ED do not compete with non-neurological patients for inpatient beds.

The neurology patients arriving at the ED possess different medical conditions based on which they can be categorized into different groups. Each patient group has different arrival rate and LOS. The patient's clinical characteristics seem to be a determinant of the extent of deterioration in health status while the patient is waiting for an inpatient bed. For example, the ED boarding time has stronger impact on health status of a severe stroke patient than that of a patient with moderate stroke (Kucukyazici et al., 2010). Under such circumstances, it may be reasonable to prioritize admission of severe stroke patients based on this observation.

These patients, however, are expected to stay longer in the ward, which may result in having more patients transferred to another hospital in the future due to the unavailability of neurology ward beds. In this context, admission policies, which prescribe the rules according to which various patients with different access time requirements are admitted to nursing wards (Hulshof et al., 2012), play an important role for improving the health outcomes of patients. The over-arching objective is to provide timely access for each emergency patient.

In designing the admission policies, the physicians face the trade-off between (i) the higher risk of deteriorated functionality due to extended ED stays for more severe patients and (ii) the increased risk of blocking due to longer length of stays of these patients. An additional trade-off is between the benefits of reducing the ED boarding time by transferring patients to another hospital and the inconvenience associated with the transfer. To address these trade-offs, an infinite-horizon average cost dynamic program (DP) is formulated and to solve large-scale problem instances an efficient approximation scheme is proposed. The objective is to minimize the

average opportunity cost of waiting and transferring by finding the most appropriate patient admission policy from the ED.

The solution approach developed in this study is based on the LP-based approximate dynamic programming (ADP). While this method typically involves solving a large-scale linear program (LP) (e.g., de Farias and Van Roy, 2006), the approach used in this study involves solving a collection of small DPs, which tends to be easier. The small DPs are derived from the LP formulation of the corresponding DP by employing a nonlinear functional approximation. The latter is informed by a static queuing approximation that results in a nonlinear programming problem capturing the uncertainties pertaining to the underlying processes. In the context of realistic problem instances based on data obtained from the MNH, the performance of the proposed admission policy is compared with that of other static policies.

The related literature has been reviewed from contextual perspective in Chapter 2. From a methodological perspective, the papers that use ADP for patient scheduling and admission problems are also relevant to this chapter. Green et al. (2006) considers capacity allocation of a diagnostic medical facility between different types of patients. The authors develop a finite-horizon DP which is approximated using linear profit functions and a heuristic policy has been generated based on this linear approximation. Patrick et al. (2008) formulates advance scheduling of patients with multiple priorities for a diagnostic facility as a discounted infinite-horizon MDP. By considering an affine linear approximation for value functions, the authors produce an approximate linear program (ALP) which is solved by applying column generation

technique on its dual problem. Using the solution of the ALP, they develop a booking policy and present the optimality gaps. The same approach has been used by Sauré et al. (2012) to schedule cancer patients for radiation therapy sessions. These types of patients require more than one appointment over the planning horizon while Patrick et al. (2008) assumes each patient requires only one appointment.

Before turning to the model statement, it is important to highlight the differentiating characteristics of this work from other studies. To the best of the author's knowledge, this is the first study that makes an explicit effort to model the differentiating features of neurology wards, and hence provides managerial insights specific to this domain. The contributions of this study are three-fold. First, from a modeling perspective, the significance of the presence of a specialized team of care providers in neurology wards is recognized, which renders off-servicing policies infeasible for neurology patients. In dealing with hard capacity constraints, the possibility of patient transfers to other hospitals is incorporated into the modeling, which is not well studied in the prevailing literature.

Second, from the viewpoint of methodology, an LP-based approximate dynamic programming (ADP) approach is developed that involves solving a number of small DPs derived by employing a non-linear functional approximation. The subsequent complexity is tackled by a novel decomposition that results in smaller DPs. An ADP-based Priority Cut-off policy is also developed that not only performs well by incorporating the state of the system in making the patient admission decisions, but also is easy to implement.

Lastly, on the managerial side, the weaknesses of the static patient admission and the ad-hoc patient transfer policies that are currently popular will be highlighted. In particular, it will be shown that by incorporating the current utilization of the ward and the nature of the waiting line, it is possible to achieve lower costs and better trade-offs between waiting times and patient transfers.

3.2. Problem Description

In this section, the problem of admitting patients with different clinical conditions into a neurology ward is described more precisely. There are n types of patients indexed by $i \in \{1, \ldots, n\}$ where type 1 is the least severe patient and type n is the most severe patient. There are B beds available in the ward. Each bed can be used for admitting the patient irrespective of her neurological condition. Patients usually wait in the ED before a bed inside the ward is assigned to them.

It is generally undesirable to keep neurology patients in the ED due to the lack of the special care needed by this group. The health status of a patient with severe condition deteriorates much faster than one with a non-severe condition, in response to delays in admission to the ward. Assuming that dis-utilities associated with such delays can be expressed in monetary terms, let π_i denote the waiting cost per unit time for a patient of type i; $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_n)$. Note that $\pi_i \leq \pi_j$ for $i < j$.

Type-i patients arrive at the system according to a Poisson process with the rate of λ_i patients per unit time. Upon the arrival of a new patient, the ward manager decides whether to accept or transfer the patient to another hospital. Transferring a type-i patient to another hospital incurs a lump-sum cost, denoted by κ_i. Let $\boldsymbol{\kappa} = (\kappa_1, \kappa_2, \ldots, \kappa_n)$. If the patient is accepted, she might be admitted to the ward

immediately unless she has to join the queue and wait until a bed becomes available for her.

Whenever a type-i patient is admitted to a bed, she occupies the bed for a time which is exponentially distributed with the mean μ_i^{-1} (which is called average LOS). Consequently, μ_i indicates the discharge rate for patients of type i. For patients with the same disease, the average LOS in more severe patients tends to be longer. The arrivals and discharges are assumed that occur independently from each other. When a patient is discharged, a decision is made on whether to admit a patient from the queue to the ward. The decision-making process should be based on the number of waiting patients from each type and also the number of beds occupied by each group of patients.

3.3. The Dynamic Program Formulation

To find the best admission policy, the problem is formulated as a continuous time dynamic program. By developing a continuous time dynamic program, it is sufficient to limit our attention only to those times when there is a change in the state of the system, while the system is being tracked at all times (Puterman, 1994). The change in the state of the system can be either an arrival of a patient or a discharge of a patient from the ward. The time horizon is considered to be infinite which is consistent with the idea of running a hospital ward. This problem can be formulated either as a total discounted cost or an average cost model. While the total discounted approach seems easier to apply, the dependency of the optimal policy on the discount factor and the initial state is a major drawback. Thus, an

average cost dynamic program will be used, where the objective is to minimize the long-run average cost of the system.

3.3.1 State Variables

The state of the system includes information about the number of waiting patients and the number of occupied beds by each patient type. It is required to distinguish between the beds occupied by different patient types because the discharge rate is not the same for all types. Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$, where x_i is the number of waiting patients of type i, and $\mathbf{b} = (b_1, b_2, \ldots, b_n)^{\mathrm{T}}$, where b_i is the number of beds occupied by patients of type i. The state of the system is given by $(\mathbf{x}, \mathbf{b})$. Note that $\mathbf{x}$ and $\mathbf{b}$ are n-dimensional column vectors. The total number of patients waiting in the system is constrained by K, which reflects the physical capacity of the ED to board neurology patients. Hence, we have $\sum_{i=1}^{n} x_i \leq K$. At any time, at most B patients are in the beds, i.e., $\sum_{i=1}^{n} b_i \leq B$. So the state space is finite. The state variables are defined as *post-action* variables so that the transition rate depends only on the state of the system and not on the actions. The following figure shows the time-line according to which the state transition occurs.

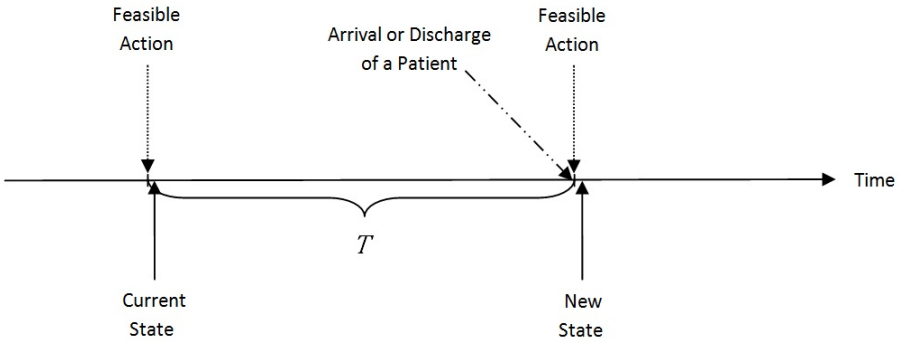

Figure 3–1: State transition in the patient admission problem

3.3.2 Actions

In a continuous time DP, the moments that a decision is made are restricted to those times that the state of the system changes (Puterman, 1994). The possible actions are classified based on the cause of state changes.

In the case of an *arrival*, the possible actions are:

- letting the patient join the queue;
- admitting the patient to the ward; and
- transferring the patient to another hospital.

The first option is not feasible if the number of waiting patients has reached its maximum capacity (K). The second option can only admit a patient if there is at least one bed available in the ward. The last option is always available.

Given state $(\mathbf{x}, \mathbf{b})$, the set of admissible actions in the case a type-i arrival is

$$\mathcal{U}_i(\mathbf{x}, \mathbf{b}) = \left\{ (a_i, t_i) \in \{0,1\}^2 \middle| a_i \leq \mathbb{I}\left\{ \sum_{j=1}^{n} b_j < B \right\}, \mathbb{I}\left\{ \sum_{j=1}^{n} x_j = K \right\} \leq a_i + t_i \leq 1 \right\} \quad (3.1)$$

where $\mathbb{I}\{\cdot\}$ is the indicator function, i.e., it takes the value of 1 if the conditions in the bracket is satisfied and is equal to 0 otherwise. The variable a_i is a 0-1 variable that represents the admission of a type-i arrival or equivalently, a type-i patient from the queue. An admission can occur only when there is at least one empty bed. The constraint $a_i \leq \mathbb{I}\left\{\sum_{j=1}^{n} b_j < B\right\}$ takes care of this issue. The variable t_i is also a 0-1 variable that indicates the decision related to transferring the new arrival. In the situation that the waiting area is full, we must either admit or transfer a patient. The constraint $\mathbb{I}\left\{\sum_{j=1}^{n} x_j = K\right\} \leq a_i + t_i \leq 1$ takes into account this requirement

when choosing an action. When $(a_i, t_i) = (0, 0)$, the patient simply joins the queue and waits until admission to the ward.

In the case of a *discharge*, the possible actions are:

- doing nothing;
- admitting one patient from the queue.

When a type-i patient is discharged, the set of feasible actions is

$$\mathcal{D}_i(\mathbf{x}) = \left\{ (d_{i1}, \ldots, d_{in}) \in \{0,1\}^n \middle| d_{ij} \leq x_j, \forall j; \sum_{j=1}^{n} d_{ij} \leq 1 \right\}. \tag{3.2}$$

The variable d_{ij} is a 0-1 variable where $d_{ij} = 1$ represents the admission of a type-j patient when a type-i patient is discharged. Obviously, this can happen only when there is at least one waiting patient of type j. The constraint $d_{ij} \leq x_j$ forces d_{ij} to the value 0 when there is no waiting patient of type j. The constraint $\sum_{j=1}^{n} d_{ij} \leq 1$ states that we can admit at most one patient from all types. When all d_{ij} are zeros, it refers to choosing not to admit any patient.

3.3.3 Transition Probabilities

Let T denote the random time between two decision points. To find the distribution of T, the following Lemma is used (Porteus, 2002).

Lemma 3.3.1 *Assume that the system is in the state* $(\mathbf{x}, \mathbf{b})$. *If the time between two arrivals of type-i patients (denoted by* T_i^a*) is distributed exponentially with parameter* λ_i *and the time between two discharges of type-i patients (denoted by* T_i^d*) is distributed exponentially with parameter* $b_i\mu_i$ *and all the arrivals and discharges are independent of each other, and* T *is the time to next state transition, then*

1. $T = \min(\min_i T_i^a, \min_i T_i^d)$ *and is exponentially distributed with parameter* $\sum_{i=1}^{n}(\lambda_i + b_i\mu_i)$,
2. $Pr(T = T_i^a) = Pr(T = T_i^a | T = t) = \frac{\lambda_i}{\sum_{i=1}^{n}(\lambda_i + b_i\mu_i)}$, *and*
3. $Pr(T = T_i^d) = Pr(T = T_i^d | T = t) = \frac{b_i\mu_i}{\sum_{i=1}^{n}(\lambda_i + b_i\mu_i)}$.

This Lemma establishes that the time to the next transition is exponentially distributed when all the events follow Poisson processes. The rate of the distribution is the sum of all rates; $\nu(\mathbf{x}, \mathbf{b}) = \sum_{i=1}^{n}(\lambda_i + b_i\mu_i)$. Also, when a transition has already happened at time t, the probability that the transition is caused by a specific event is the rate of that event divided by the sum of all rates. This probability is independent of the time that has passed. Since the state of the system changes over time, the transition rate in each period is not constant. To transform the system into a Markov chain with uniform transition rate, the *uniformization* technique is applied.

To use the *uniformization* technique, note that an upper bound for the transition rate is $\nu^{\max} = \sum_i \lambda_i + b\mu^{\max}$ where $\mu^{\max} = \max_i \mu_i$. So the new transition probabilities are given as follows (Bertsekas, 2005):

$$\text{Transition Probability} = \begin{cases} \frac{\lambda_i}{\nu^{\max}}, & \text{if there is an arrival of type } i, \\ \frac{b_i\mu_i}{\nu^{\max}}, & \text{if there is a discharge of type } i, \\ 1 - \frac{\sum_{i=1}^{n}(\lambda_i + b_i\mu_i)}{\nu^{\max}}, & \text{if there is no change in state.} \end{cases}$$

The time is scaled such that the maximum transition rate ($\nu^{\max}$) is normalized to 1. To do so, the new arrival and service rates are defined as: $\lambda_i^{'} = \frac{\lambda_i}{\nu^{\max}}$ and $\mu_i^{'} = \frac{\mu_i}{\nu^{\max}}$, for all i. Then the new transition probabilities are:

$$\text{(Normalized) Transition Probability} = \begin{cases} \lambda_i^{'}, & \text{if there is an arrival of type } i, \\ b_i\mu_i^{'}, & \text{if there is a discharge of type } i, \\ 1-\sum_{i=1}^{n}(\lambda_i^{'}+b_i\mu_i^{'}), & \text{if there is no} \\ & \text{change in state.} \end{cases}$$

Accordingly, the waiting cost of type-i patients per each normalized time interval is $\pi_i^{'} = \frac{\pi_i}{\nu^{\max}}$. For notational simplicity, λ_i, μ_i, and $\boldsymbol{\pi}$ denote the normalized parameters in the remainder of this chapter.

3.3.4 Bellman Optimality Equation

The Bellman equation of dynamic program is given by

$$\begin{aligned} (\mathbf{DP}) \quad h(\mathbf{x},\mathbf{b}) = \boldsymbol{\pi}^T\mathbf{x} - \rho^* &+ \sum_{i=1}^{n}\lambda_i \min_{\mathbf{a}_i\in\mathcal{U}_i(\mathbf{x},\mathbf{b})}\{\kappa_i t_i + h(\mathbf{x}+(1-a_i-t_i)\mathbf{e}_i, \mathbf{b}+a_i\mathbf{e}_i)\} \\ &+ \sum_{i=1}^{n} b_i\mu_i \min_{\mathbf{d}_i\in\mathcal{D}_i(\mathbf{x},\mathbf{b})}\{h(\mathbf{x}-\mathbf{d}_i, \mathbf{b}-\mathbf{e}_i+\mathbf{d}_i)\} \\ &+ \left(1-\sum_{i=1}^{n}\lambda_i - \sum_{i=1}^{n} b_i\mu_i\right) h(\mathbf{x},\mathbf{b}), \qquad \forall \mathbf{x},\mathbf{b}, \end{aligned}$$

where $\mathcal{U}_i(\mathbf{x},\mathbf{b})$ and $\mathcal{D}_i(\mathbf{x})$ are given by (3.1) and (3.2), respectively, and $\mathbf{e}_i$ is an n-dimensional identity vector.

In the (**DP**), ρ^* is the optimal average cost per normalized time period and $h(\mathbf{x},\mathbf{b})$ is the bias function which represents the total difference from optimal average cost over all periods if we start from state $(\mathbf{x},\mathbf{b})$. The term $\boldsymbol{\pi}^T\mathbf{x}-\rho^*$ is the difference between the cost of this period and the optimal average cost. The second (third) term refers to the case when a type-i patient arrives (discharges). The last term is

associated with the case of no change at the end of period and thus we are returning to the same state. This term has been added due to the *uniformization*. In this model, all the states can be reached from other states, i.e. the *Weak Accessibility* holds (Bertsekas, 2005). Thus, the optimal average cost is independent of the initial state of the system.

3.4. The Data

In order to demonstrate the applicability of the proposed DP formulation and to garner managerial insights, a full data set representing the patient flows through the 3-South neurology ward of the MNH has been developed. As mentioned before, the MNH neurology ward has sixteen inpatient beds. In an effort to focus on the care provided to stroke patients, the patients are categorized into four patient types: Mild Non-Stroke, Mild Stroke, Severe Non-Stroke, and Severe Stroke. Note that these patients arrive at the MNH through the ED and are kept boarding there until a bed becomes available at the ward.

This data set includes all the patients treated in the neurology ward for three full fiscal years. The sources of this data set are: the hospital's ED information system, the patient registry of McGill University Health Center, and the paper-based patient charts of the stroke and non-stroke patients admitted to the 3-South neurology ward of the MNH. In this section, the assumptions regarding the arrival and LOS distributions are verified using the available data. Furthermore, the waiting and transfer costs in the model are estimated in the form of health related quality of life (HRQoL) (Xie et al., 2006).

3.4.1 Patient Arrival and LOS Distributions

The patient inter-arrival times to the system, and the number of patient departures during a given time interval from the hospital, are random and dependent on the patient category. The parameters of the arrival and departure distributions are based on actual historical data. Based on the χ^2 test results of Tables 3–1 and 3–2 the Poisson arrival and exponential LOS hypotheses cannot be rejected.

Patient Type	Arrival Process (daily)		
	Mean (λ)	Variance (σ^2)	H_0^* = Arrival is Poisson
Mild Non-Stroke	0.236	0.246	Not Rejected ($0.05 < p < 0.10$)
Mild Stroke	0.262	0.291	Not Rejected ($0.10 < p < 0.25$)
Severe Non-Stroke	0.139	0.141	Not Rejected ($0.25 < p < 0.50$)
Severe Stroke	0.113	0.117	Not Rejected ($0.25 < p < 0.50$)

Table 3–1: Goodness-of-fit tests on arrival process of all patient types (* All tests are one-tailed with $\alpha = 0.05$)

The seasonal variation in the arrival pattern of stroke and non-stroke neurological patients, i.e., the impact of seasonality in the number of arrivals for each patient type, is also investigated. To this end, the cumulative number of arrivals for each patient type by month over a time period of three years is calculated and the term of *seasonality* refers to monthly variation over cumulative number of arrivals. For modeling the seasonal variation, the cosine function is used, which is a simple curve of cyclic periodicity. The estimated monthly number of arrivals with the best-fit cosine curve function, the amplitude of the seasonality, and the Roger's R values (Roger,

Patient Type	LOS (day)		
	Mean ($\frac{1}{\mu}$)	Variance (σ^2)	H_0^* = LOS is exponential
Mild Non-Stroke	13.003	162.383	Not Rejected ($0.25 < p < 0.50$)
Mild Stroke	11.491	114.204	Not Rejected ($0.05 < p < 0.10$)
Severe Non-Stroke	19.011	305.904	Not Rejected ($0.10 < p < 0.25$)
Severe Stroke	22.002	596.445	Not Rejected ($0.05 < p < 0.10$)

Table 3–2: Goodness-of-fit tests on LOS distribution of all patient types (* All tests are one-tailed with $\alpha = 0.05$)

1977) are calculated in order to examine the significance of seasonal variation. The results in Table 3–3 indicate that seasonality is not significant for all patient types.

Patient Type	Monthly Mean	Amplitude	R^2
Mild Non-Stroke	21.5	0.9*	0.12
Mild Stroke	24.0	0.5*	0.08
Severe Non-Stroke	12.5	1.1*	0.25
Severe Stroke	10.5	1.04*	0. 17

Table 3–3: Seasonal variation in arrival of all patient types (* p > 0.05)

The details of the cumulative number of monthly arrivals are provided in Figure 3–2, where the figure does not visually display any consistent seasonal patterns in the arrival numbers.

3.4.2 Patients Waiting and Transfer Costs

Waiting Cost: As mentioned earlier, the patients boarding in the ED for a bed in the neurology ward suffer from lack of specialized care and their health status deteriorates as a consequence of staying in the ED. This deterioration emerges as worse

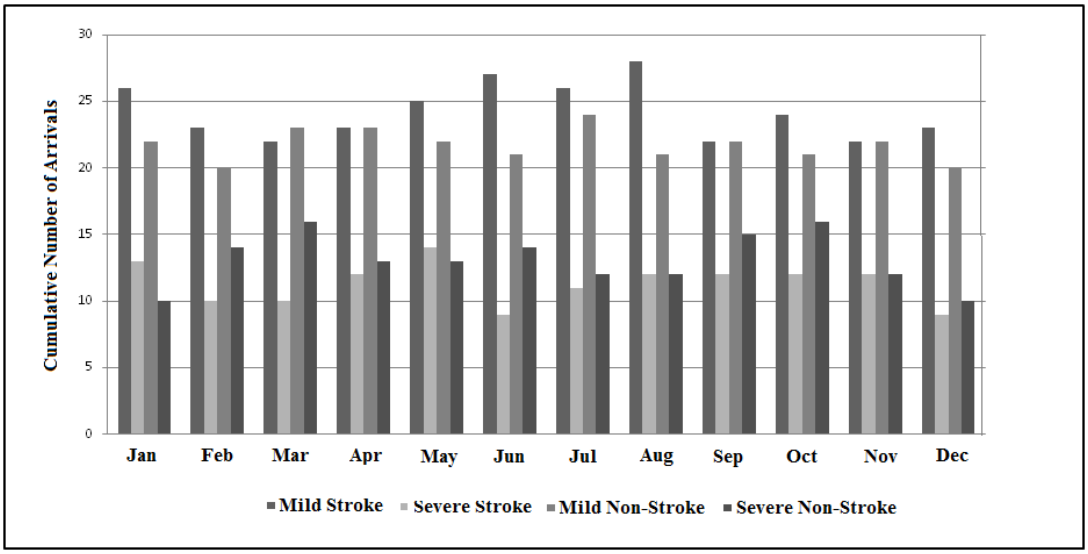

Figure 3–2: Cumulative number of monthly arrivals for each patient type over three years

functionality of the patients, which is one of the most important health outcomes of the neurological patients. For those patients, discharge destination can be used as a proxy of patient's functionality at the time of discharge. In this context, Kucukyazici et al. (2010) have found that longer ED boarding time is strongly associated with increased probability of not being able to discharge to home, i.e., being admitted to rehabilitation center or long term care facility. To be more specific, they observed that a 10% increase in the ED LOS is related to a 7.7% increase in the risk of being discharged to either a rehabilitation center or a long term care facility, i.e., not being able to go home. It is established that the discharge destination has a significant impact on both short-term and long-term HRQoL (Xie et al., 2006). Thus, the patient's waiting cost can be estimated as the expected HRQoL lost resulting from not being able to go home due to the ED boarding.

Let β_i denote the increase in the probability of not being discharged home for type-i patients, as a result of one time unit boarding in the ED. The patient type specific β_i are estimated utilizing a regression model controlled for all other clinical and demographic factors.

Let s_i^{R} and s_i^{L} denote the conditional probabilities of being sent to a rehabilitation center and a long term care facility, respectively; given that the type-i patient is not discharged to home. Note that $s_i^{\mathrm{R}} + s_i^{\mathrm{L}} = 1$. Moreover, the HRQoL are defined as the values associated with discharge destination as $\mathrm{Q_H}$, $\mathrm{Q_R}$, and $\mathrm{Q_L}$ for home, rehabilitation center, and long term care facility, respectively.

There are several studies in the literature that report the HRQoL measures for neurological patients including Hopman and Verner (2003), Jaracz and Kozubski (2003), Jönsson et al. (2005), and Nichols-Larsen et al. (2005). In this model, the short-term HRQoL measures estimated by Nichols-Larsen et al. (2005) are used and the HRQoL of being discharged to home is normalized to 100.

The waiting cost per unit time for type-i patient, π_i, is the expected loss in quality of health outcomes as a consequence of one unit time increase in the ED boarding time:

$$\pi_i = \beta_i p_i \left(\mathrm{Q_H} - (s_i^{\mathrm{R}} \mathrm{Q_R} + s_i^{\mathrm{L}} \mathrm{Q_L}) \right), \tag{3.3}$$

where p_i corresponds to the average probability of discharge to rehabilitation center and long term care facility of patient type-i for the group of patients who do not experience any delay in the ED. Using Equation (3.3), the waiting cost per day for each patient type is estimated and presented in Table 3–4.

Index (i)	Patient Type	Daily Waiting Cost (π_i)
1	Mild Non-Stroke	70
2	Mild Stroke	90
3	Severe Non-Stroke	145
4	Severe Stroke	295

Table 3–4: Estimated waiting cost of neurology patients in the ED per day

Transfer Cost: The existing clinical guidelines used at the MNH recommend to transfer the patients to another hospital if their waiting time in the ED exceeds 48 hours. This means that the ward manager is willing to keep the patients in the ED for two days and if no bed becomes available in that period the patient is transferred to another hospital, where the patient is presumably admitted to the ward without any delay. Kucukyazici (2010) studies the process of patient transfer to other hospital by means of a comprehensive simulation model of MNH ED, neuro-ICU, and neurology ward. The results in that study clearly demonstrate that the current practice of waiting for 48 hours of ED boarding until a transfer decision is made, is not the best policy. Thus, the model proposed in this study assumes that the transfer decision can be made at the time of patient arrival based on the overall congestion of the system. Consequently, if we decide to transfer the patient, the maximum transfer cost is considered to be equivalent to two days of waiting in the current hospital's ED. In general, if the threshold for transferring type-i patients in a hospital is d_i time units and π_i is the ED waiting cost per unit time, then the transfer cost for type-i patient is estimated as:

$$\kappa_i = d_i \pi_i. \tag{3.4}$$

3.5. Properties of the Optimal Policy

In this section, some properties of the optimal policy are illustrated by means of a special case of the problem with two types of patients, mild stroke (referred to as type 1) and severe stroke (referred to as type 2). The arrival rate and average LOS for these two types can be found in Tables 3–1 and 3–2. For this special case of problem, the number of available beds and the ED capacity are assumed to be equal to eight (half of the current bed capacity at the MNH), i.e., $B = K = 8$. Through solving a large number of problem instances, the structure of the optimal admission policy has been explored. In the remainder of this section, only the most revealing instances are reported that, in turn, are grouped into two subsections – corresponding to the arrival and the discharge of a stroke patient, respectively.

As observed in the following examples, the form of the optimal policy does not seem to be straightforward. The complexity of the problem requires that the optimal policy be based on all the information about the system. In particular, it is not sufficient to know how many beds are occupied (or equivalently, how many beds are available). Instead, we need to know the number of occupied beds and the number of people waiting by each patient type to make the best decision regarding the admission of a new patient. It also seems that the optimal policy is robust with respect to the magnitude of waiting costs and is affected only by their ratio.

3.5.1 Arrival of a Stroke Patient

Illustrative Example I: The starting cost parameters are: $\boldsymbol{\pi} = (90, 295)$ as in Table 3–4, and $\boldsymbol{\kappa} = 2\boldsymbol{\pi}$. We first assume that there is a new arrival of a mild stroke patient in the system while $x_2 = 0$ and $b_2 = 0$ (i.e., no severe stroke patient

is either in the queue or in the ward). The optimal decision in such a situation is to admit the new arrival to the ward if there is an available bed, and transfer the patient otherwise. This implies that there is no tendency to reserve a bed for a severe patient that might arrive in the future. If we increase the waiting cost of severe stroke patients, however, a different behavior is observed. For example, when $\pi_2 = 5\pi_1$, $x_1 \leq 5$, and if there is only one available bed, the optimal policy is to keep that bed for the possible arrivals of severe patients.

Illustrative Example II: The condition $x_1 \leq 5$ in the above example implies a threshold policy. Denote the threshold on the number of mild patients waiting for a bed by τ. Note that $\tau = 5$ in that case. To see the impact of the number of different patient types in the ward on τ, a parametric analysis is conducted on b_1 while fixing $b_1 + b_2 = B - 1$ (i.e., there is always one available bed). Let $\pi_2 = 5\pi_1$ and $x_2 = 0$. The optimal policy in the case of an arrival of mild stroke patient is shown in Figure 3–3. When $b_1 \leq 2$, we reserve the available bed for a severe patient if $x_1 \leq 3$ ($\tau = 3$) and allocate that bed to the mild stroke patients, otherwise. This threshold increases by one, i.e., $\tau = 4$, when $3 \leq b_1 \leq 4$, and increases by two, i.e., $\tau = 5$, when $5 \leq b_1 \leq 7$. *Observe that the reservation threshold on the number of waiting patients (up to which we earmark a bed for a severe stroke patient) increases when the number of beds occupied by the mild stroke patients increases.* This sounds counter-intuitive as we expect this threshold to decrease. When more beds are occupied by mild patients, it is more likely to have an empty bed in the near future. Therefore, we might be better off letting the mild patients use the only available bed sooner. Nevertheless, the optimal policy suggests an opposite behavior.

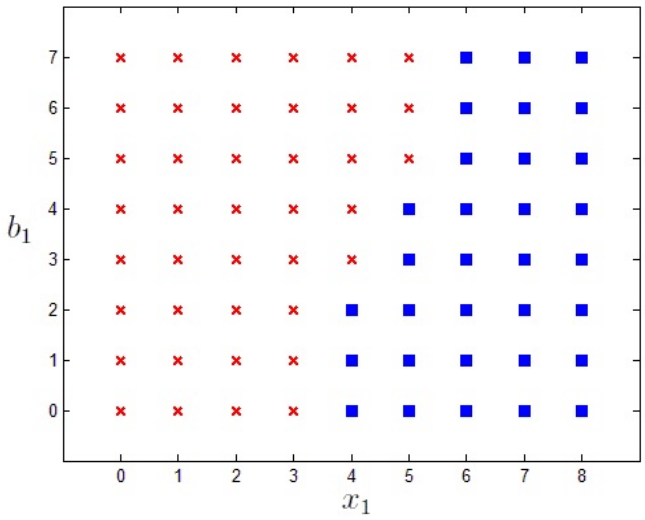

Figure 3–3: Optimal decision if a mild stroke patient arrives while $x_2 = 0$ and $b_1 + b_2 = B - 1$
$\times$ = Transfer (Reservation) ■ = Admit mild stroke patient to bed

Illustrative Example III: In Figure 3–3, it is interesting to note that when there are less than four mild stroke patients waiting, the new arrival of this type will be transferred regardless of the value of b_1. Therefore, it is unlikely for the system to reach the state where $x_1 > 4$ and $x_2 = 0$. In order to study the more likely system states, let us consider system states in which $x_1 = x_2 = 0$. Let e be the number of empty beds. For all values of e, the optimal policy in the event of a mild patient arrival is shown in Figure 3–4. Note that in this Figure the waiting cost of severe patients has been decreased to 325 from 450 in the previous example. The optimal policy is to admit the mild patient to the bed as long as more than two beds are available. If no bed is available, the patient is transferred. However, if only one bed is free, we look at the patient mix in the ward. As opposed to the previous example,

more beds being occupied by mild patients supports admitting the mild patient to the bed.

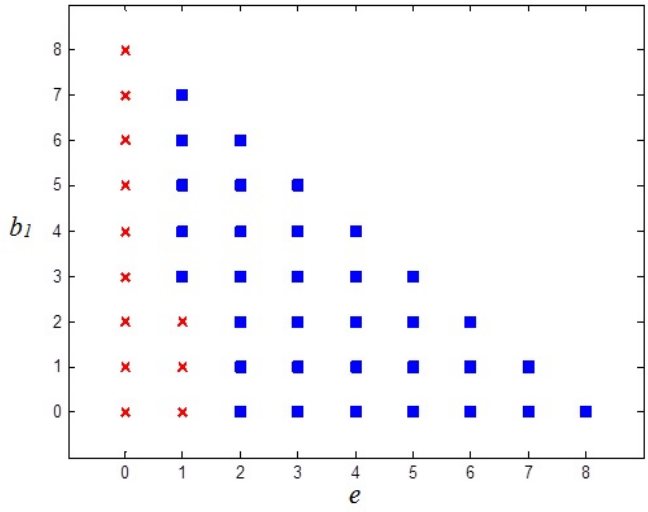

Figure 3–4: Optimal decision if a mild stroke patient arrives while $x_1 = x_2 = 0$, $b_1 + b_2 = B - e$
$\times$ = Transfer (Reservation) ■ = Admit mild stroke patient to bed

Illustrative Example IV: Now let us consider the arrival of a severe stroke patient to the system. When $\boldsymbol{\pi} = (90, 295)$, the optimal policy recommends that the severe patient is admitted to the bed in most cases. If the cost parameters are set to $\boldsymbol{\pi} = (90, 135)$, i.e., $\pi_2 = 1.5\pi_1$, we see some transfers of severe patients . If there is one bed available, all other beds are occupied by severe patients, and $x_1 \geq 7$, then we are better off transferring the newly arrived severe patient to another hospital and using the available bed for mild stroke patients.

3.5.2 Discharge of a Stroke Patient

Illustrative Example V: Assume that all beds are occupied and a patient is discharged, i.e., one bed becomes available. The cost parameters are first set at $\boldsymbol{\pi} = (200, 250)$, and $\boldsymbol{\kappa} = 2\boldsymbol{\pi}$. Moreover, an arbitrary combination of occupied beds is chosen; for example, $b_1 = b_2 = B/2$. The optimal decision when a mild stroke patient is discharged for all combinations of x_1 and x_2 is shown in Figure 3–5. In this figure, it is interesting to observe that when $1 \leq x_1 \leq 3$ and one bed becomes available, we begin by assigning that bed to a severe stroke patient. But if the number of severe stroke patients increases, it would be better to give that bed to a mild stroke patient, which seems to be counter-intuitive. Why should we switch from prioritizing severe stroke patients to letting a mild stroke patient occupy the bed when we have more severe patients in the queue? The reason is the slower discharge rate of severe stroke patients. As the system gets more congested, a tendency is formed toward serving the mild stroke patients who have higher discharge rate and hence a higher chance of emptying that bed in the near future. This phenomenon can happen when the waiting costs for the two types are close to each other. If we increase the waiting cost for the severe patients, this phenomenon disappears.

Illustrative Example VI: We might also be interested in seeking cases in which we reserve a newly freed bed for a future arrival of severe stroke patients. To examine this, the cost parameters are changed to $\boldsymbol{\pi} = (100, 1500)$, and $\boldsymbol{\kappa} = 2\boldsymbol{\pi}$. Let us assume $x_2 = 0$, otherwise bed reservation for severe stroke patients cannot be optimal. In addition, all beds are occupied (i.e., $b_1 + b_2 = B$) when a mild stroke patient is discharged. To demonstrate how the reservation pattern changes

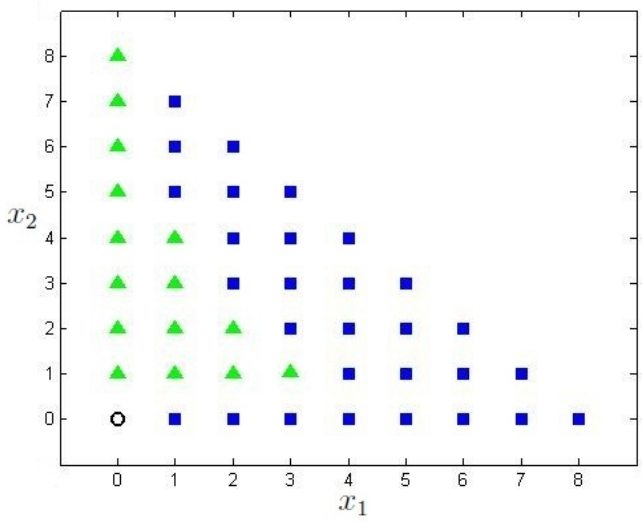

Figure 3–5: Optimal decision if a mild stroke patient is discharged while $b_1 = b_2 = 4$. $\circ$ = No action ■ = Admit mild stroke patient to bed ▲ = Admit severe stroke patient to bed

depending on the patient population in the ward b_1 is varied from 1 to B. Figure 3–6 depicts the optimal decision in this parametric analysis. Evidently, it is optimal to reserve a bed for severe stroke patients when their waiting cost is much higher than that of the mild stroke patients. Note that the threshold on x_1 above which we stop reserving increases as the number of occupied beds by mild stroke patients (b_1) increases. This is similar to the counter-intuitive observation in Figure 3–3.

Remarks: The illustrative examples demonstrate the complexity in the structure of the optimal policy. Even though these examples suggest some special forms of admission policy (threshold policy), the precise form of the optimal policy is quite intricate and varies with the model parameters.

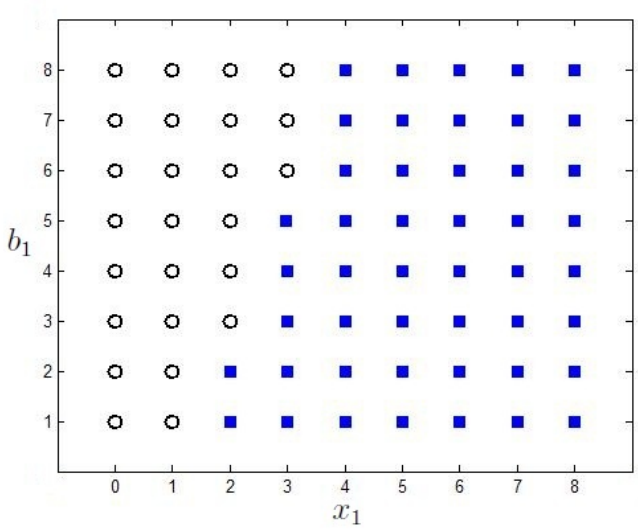

Figure 3–6: Optimal decision if a mild stroke patient is discharged while $x_2 = 0$ and $b_1 + b_2 = 8$.
$\circ$ = No action (Reservation) ■ = Admit mild stroke patient to bed

3.6. The Solution Methodology

The Bellman equation for an average cost DP can be solved with the relative value iteration algorithm in a reasonable amount of time when the size of the problem is relatively small. As the number of patient types, the number of beds, or the waiting room capacity increases, the *curse of dimensionality* hinders us from a brute-force solution of the DP. Thus, an approximation scheme is proposed to find a good admission policy in large-scale instances of the problem.

The proposed approach involves two steps. The first step is to build a static model in which we assume the beds are allocated to different patient types and the allocation does not change over time. By solving this static model, we find the number of beds that should be allocated to each type so that the average cost per

period is minimized. Further, the proportion of patients from each type that are transferred to another hospital is also determined. The average cost of such a model accounts for waiting costs of patients as well as transfer costs. This model is called static because the policy is fixed over time irrespective of the system state.

The second step is to develop an approximate dynamic program (ADP) that can be solved in a reasonable time frame. To do so, some information including the opportunity cost of occupying a bed, the number of beds allocated to each type, the average waiting time, and the average queue length of each type is exploited from the static model solution. Then, using this information, the bias function $h(\mathbf{x}, \mathbf{b})$ is approximated. To be more precise, it is assumed that the bias function is the sum of contributions from all patient types. The contribution of all patient types except one type is estimated using a non-linear function and for this specific type we leave its contribution unknown. This approximate bias function is plugged back into the Bellman equation that ultimately leads to a simpler DP to solve. In the resulting DP, we will deal with only one type of patient (that specific type for which the contribution is left unknown). We iterate this procedure for all types of patients. In the end, we sum all the contributions up to approximate the bias function; $h(\mathbf{x}, \mathbf{b})$. Based on the approximated bias function, we can create an admission policy. Figure 3–7 shows all these steps and their interactions.

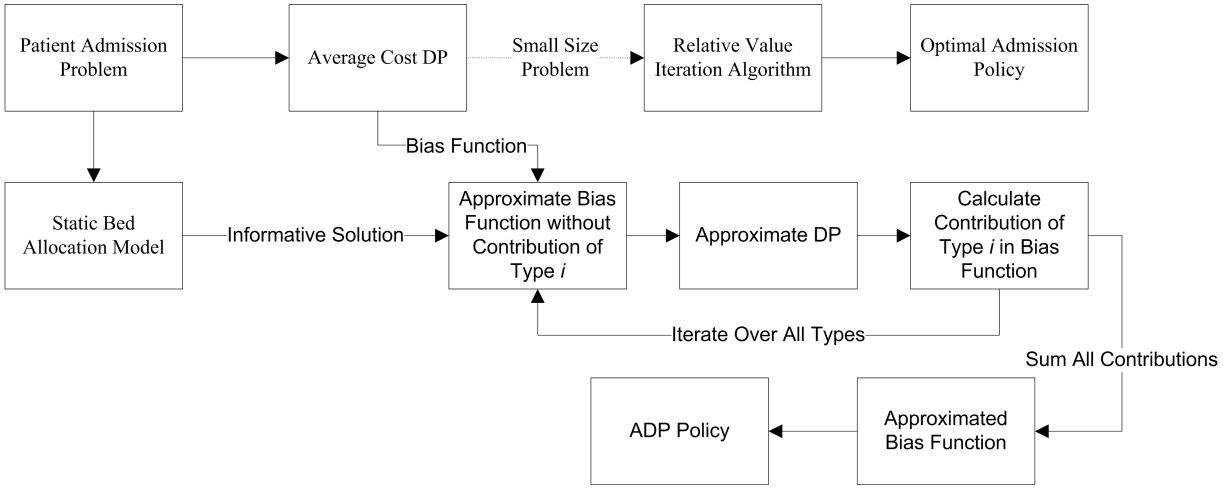

Figure 3–7: Schematic view of the solution methodology

3.6.1 The Static Model

This section presents a static model that is based on queueing approximation of the problem. This static model allocates a certain number of beds exclusively for each type of patients. As opposed to the dynamic optimal policy obtained from the Bellman equation, this model determines a static policy which does not change over time and is not influenced by the state of the system.

Suppose the number of beds dedicated to type-i patients is $\tilde{b}_i$. The system with $\tilde{b}_i$ beds serving incoming type-i patients can be viewed as a queue with $\tilde{b}_i$ servers. Due to the constraint on the total number of waiting patients, the type of queue we are dealing with for type-i patients is an $M/M/\tilde{b}_i/\tilde{b}_i + k_i$ queue. Here k_i is the upper bound on the length of queue for type-i patients, above which the new arrivals will be turned away. The service rate is μ_i but the arrival rate should not necessarily be equal to λ_i, because we can transfer some patients upon their arrival to another

hospital. So the rate of patients entering the system can be less than the original arrival rate. Therefore, the decision variable $\tilde{\lambda}_i$ is defined as the adjusted arrival rate.

The total average cost of this queue is the sum of the average waiting cost of the patients and the average cost of transferring the new arrivals. Let us denote the average number of waiting patients of type i in the queue by L_i. So the average waiting cost is given by L_i times the waiting cost per unit time. Also, on average, $(\lambda_i - \tilde{\lambda}_i)$ of type-i patients are transferred to another hospital per unit time. Note that a portion of new arrivals will be blocked due to lack of space in the waiting area, which is $\tilde{\lambda}_i p_{k_i}$ (p_{k_i} is the probability that there are k_i patients waiting in the queue). So in total, $\lambda_i - \tilde{\lambda}_i(1 - p_{k_i})$ of the arrivals are transferred. The associated transfer cost would be $\kappa_i \left(\lambda_i - \tilde{\lambda}_i(1 - p_{k_i})\right)$ per unit time.

In a general $M/M/c/c+k$ queue, with arrival rate of λ and service rate of μ, the average length of queue is given by (Gross et al., 2008)

$$L = \begin{cases} \frac{p_o r^c \rho}{c!(1-\rho)^2}[1 - \rho^{k+1} - (1-\rho)(k+1)\rho^k], & (\rho \neq 1)\,, \\ \frac{p_o r^c}{c!}\frac{k(k+1)}{2}, & (\rho = 1)\,, \end{cases} \tag{3.5}$$

where $r = \lambda/\mu$ and $\rho = r/c$. The blocking probability is calculated using

$$p_k = \frac{r^{c+k}}{c! c^k} p_o, \tag{3.6}$$

where

$$p_o = \begin{cases} \left[\frac{r^c}{c!}(\frac{1-\rho^{k+1}}{1-\rho}) + \sum_{n=0}^{c-1} \frac{r^n}{n!}\right]^{-1}, & (\rho \neq 1)\,, \\ \left[\frac{r^c}{c!}(k+1) + \sum_{n=0}^{c-1} \frac{r^n}{n!}\right]^{-1}, & (\rho = 1)\,. \end{cases} \tag{3.7}$$

Note that the average waiting time is obtained by $W = \frac{L}{\lambda(1-p_k)}$.

The goal of the static model is to allocate all available beds (B) and waiting room capacity (K) among different types of patients such that the average cost of the system is minimized. This can be done using the following mixed-integer program:

$$
\begin{aligned}
(\mathbf{SM}) \quad F^* = \text{Minimize} \quad & \sum_{i=1}^{n} \pi_i L_i + \sum_{i=1}^{n} \kappa_i \left(\lambda_i - \tilde{\lambda}_i (1 - p_{k_i}) \right) \\
\text{Subject to} \quad & \sum_{i=1}^{n} \tilde{b}_i \leq B, \\
& \sum_{i=1}^{n} k_i \leq K, \\
& \tilde{\lambda}_i \leq \lambda_i, \quad \forall i, \\
& \tilde{\lambda}_i, \tilde{b}_i \geq 0, \\
& \tilde{b}_i, k_i \text{ integer}, \quad \forall i.
\end{aligned}
$$

Proposition 3.6.1 *The optimal solution of the (**SM**) gives an upper bound on the optimal average cost in the (**DP**); i.e., $F^* \geq \rho^*$.*

Proof The proof of Proposition 3.6.1 is straightforward since the optimal solution of the static model is always a feasible policy for (**DP**).

In order to solve the static model as a continuous non-linear program, the integrality constraints on the number of allocated beds ($\tilde{b}_i$) and waiting room capacity (k_i) are relaxed. To find the length of queue when the number of beds is not integer, the following algorithm can be used. It also provides the values for blocking probabilities.

1. If $\tilde{\lambda}_i = 0$, then $L_i = 0$ and $p_{k_i} = 0$.
2. If $\tilde{\lambda}_i \neq 0$ and $\tilde{b}_i = 0$, then $L_i = \infty$ and $p_{k_i} = 1$.

3. If $\tilde{\lambda}_i \neq 0$, $\tilde{b}_i \neq 0$ and $\tilde{b}_i$ is integer, then L_i and p_{k_i} are calculated through Equations (3.5) and (3.6).
4. If $\tilde{\lambda}_i \neq 0$ and $\tilde{b}_i \neq 0$ and $\tilde{b}_i$ is non-integer, then $\tilde{b}_i$ is rounded to nearest integer (called b_{new}) and service rate is adjusted to $\mu_{new} = \frac{\tilde{b}_i \mu_i}{b_{new}}$. The L_i and p_{k_i} are calculated using b_{new} and μ_{new}.

For non-integer values of k_i, we take the following interpolation approach:

1. $L_i(k_i) = (k_i - \lfloor k_i \rfloor)L_i(\lceil k_i \rceil) + (\lceil k_i \rceil - k_i)L_i(\lfloor k_i \rfloor)$.
2. $p_{k_i} = (k_i - \lfloor k_i \rfloor)p_{\lceil k_i \rceil} + (\lceil k_i \rceil - k_i)p_{\lfloor k_i \rfloor}$.

where $\lfloor k_i \rfloor$ and $\lceil k_i \rceil$ refer to the biggest integer number less than or equal to k_i and smallest integer number greater than or equal to k_i, respectively.

Denote the solution of (**SM**) by $(\tilde{\lambda}_i^*, \tilde{b}_i^*, k_i^*)$ for all i. Based on this solution, the maximum number of beds occupied by type-i patients is $\tilde{b}_i^*$. The number of waiting patients of type i is limited to k_i^*. Also, we reject a fraction of new arrivals so that the actual rate of patients who enter the system is $\tilde{\lambda}_i^*$. The other piece of information that is extracted from the solution of static model is the value of dual variable of the first constraint (the constraint on the number of allocated beds). The value of this variable (which we denote by α) gives how much the average cost of the system can be reduced if we have one more bed available. Therefore, it can be interpreted as the opportunity cost of occupying a bed for one unit time (or simply, value of a bed). This information will be used in deriving the approximate dynamic program and developing two static policies to use as benchmarks in computational experiments.

3.6.2 The Approximate Dynamic Programming

The (**DP**) formulation can be written as a linear program as follows:

$$
\begin{aligned}
(\mathbf{LP}) \qquad & \rho^* = \max \rho \\
& h(\mathbf{x},\mathbf{b}) + \rho \le \boldsymbol{\pi}^T \mathbf{x} + \sum_{i=1}^{n} \lambda_i \min_{\mathbf{a}_i \in \mathcal{U}_i(\mathbf{x},\mathbf{b})} \left\{ \kappa_i t_i + h(\mathbf{x} + (1 - a_i - t_i)\mathbf{e}_i, \mathbf{b} + a_i \mathbf{e}_i) \right\} \\
& \qquad + \sum_{i=1}^{n} b_i \mu_i \min_{\mathbf{d}_i \in \mathcal{D}_i(\mathbf{x})} \left\{ h(\mathbf{x} - \mathbf{d}_i, \mathbf{b} - \mathbf{e}_i + \mathbf{d}_i) \right\} \\
& \qquad + \left(1 - \sum_{i=1}^{n} \lambda_i - \sum_{i=1}^{n} b_i \mu_i \right) h(\mathbf{x},\mathbf{b}), \qquad \forall \mathbf{x}, \mathbf{b}.
\end{aligned}
$$

In the above, the decision variables are ρ and $h(\cdot)$. Note that the terms on the right hand side of the constraint can be linearized by expanding the constraint. We prefer the current, non-linear, form for development later in this section.

Recall from the solution of the (**SM**) that for type-i patients, the number of allocated beds is $\tilde{b}_i^*$, the adjusted arrival rate is $\tilde{\lambda}_i^*$, and the maximum length of queue is k_i^*. Furthermore, we can calculate the average number of type-i patients in the queue (denoted by ${L_i}^*$), and their average waiting time (denoted by ${W_i}^*$). Another piece of information that is used from the static model solution is the dual variable associated with the first constraint in the (**SM**). The value of this dual variable, as mentioned earlier, is denoted by α and is interpreted as the opportunity cost of occupying one bed per unit time.

The bias function $h(\mathbf{x},\mathbf{b})$ in the (**LP**) can be approximated by

$$
h(\mathbf{x},\mathbf{b}) \approx h_i(x_i, b_i) + \sum_{j \neq i} \left(\pi_j w_j (x_j - {L_j}^*)^+ + \frac{\alpha (b_j - \tilde{b}_j^*)^+}{\mu_j} \right), \quad \forall i, \tag{3.8}
$$

where $(y)^+ = \max(0, y)$. For each type $j \neq i$, the contribution to the bias function is estimated by $\pi_j {W_j}^*(x_j - L_j^*)^+ + \frac{\alpha(b_j - \tilde{b}_j^*)^+}{\mu_j}$ and for type i, the contribution is represented by a general function $h_i(x_i, b_i)$. For type-j patients, if we control the system according to the solution of the (**SM**), we expect to see, on average, L_j^* patients waiting in the system. So if the number of waiting patients is less than or equal to L_j^*, there is no extra cost than the average cost and the contribution is zero. But if $x_j \geq L_j^*$, then bias from the average cost can be estimated by the waiting cost of excess patients $(x_j - L_j^*)^+$. We know that from the (**SM**), a typical patient of type j is expected to wait ${W_j}^*$ units of time and the waiting cost per unit time is π_i. So the estimated contribution of the extra patients of type j is $\pi_j {W_j}^*(x_j - L_j^*)^+$.

Similarly, the cost of occupying the bed by type-j patients is estimated. For each bed occupied in addition to the allocated beds in solution of the (**SM**), $\tilde{b}_i^*$, the opportunity cost per unit time is $\alpha(b_j - \tilde{b}_j^*)^+$. We know that on average, a typical patient of type j stays in bed for μ_j^{-1} units of time. Therefore, the total opportunity cost can be expressed by $\alpha\mu_j^{-1}(b_j - \tilde{b}_j^*)^+$.

Plugging (3.8) into the (**LP**) and simplifying, we obtain a new linear program:

$$
\begin{aligned}
(\mathbf{LP1}) \quad & \max \quad \rho \\
& \text{s.t.} \quad h_i(x_i, b_i) + \rho \leq \boldsymbol{\pi}^T \mathbf{x} + \lambda_i \min_{\mathbf{a}_i \in \mathcal{U}_i(\mathbf{x},\mathbf{b})} \left\{ \kappa_i t_i + h_i(x_i + 1 - a_i - t_i, b_i + a_i) \right\} \\
& \qquad + \sum_{k \neq i} \lambda_k \min_{\mathbf{a}_k \in \mathcal{U}_k(\mathbf{x},\mathbf{b})} \left\{ \kappa_k t_k + \pi_k {W_k}^* \mathbb{I}\left\{ a_k + t_k = 0, x_k \geq L_k^* \right\} + \frac{\alpha}{\mu_k} \mathbb{I}\left\{ a_k = 1, b_k \geq \tilde{b}_k^* \right\} \right\} \\
& \qquad + b_i \mu_i \min_{\mathbf{d}_i \in \mathcal{D}_i(\mathbf{x})} \left\{ h_i(x_i - d_{ii}, b_i + d_{ii} - 1) + \sum_{j \neq i} \left(\frac{\alpha}{\mu_j} \mathbb{I}\left\{ d_{ij} = 1, b_j \geq \tilde{b}_j^* \right\} - \right. \right.
\end{aligned}
$$

$$\pi_j {W_j}^* \mathbb{I}\left\{d_{ij}=1, x_j \geq L_j^*+1\right\})\Bigg\} + \sum_{k\neq i} b_k \mu_k \min_{\mathbf{d}_k \in \mathcal{D}_k(\mathbf{x})} \Bigg\{ h_i(x_i - d_{ki}, b_i + d_{ki}) -$$

$$h_i(x_i, b_i) - \pi_k {W_k}^* \mathbb{I}\left\{d_{ki}=1, x_k \geq L_k^*+1\right\} - \frac{\alpha}{\mu_k}\mathbb{I}\left\{d_{kk}=0, b_k \geq \tilde{b}_k^*+1\right\} +$$

$$\sum_{j\neq i,k}\left(\frac{\alpha}{\mu_j}\mathbb{I}\left\{d_{kj}=1, b_j \geq \tilde{b}_j^*\right\} - \pi_j {W_j}^* \mathbb{I}\left\{d_{kj}=1, x_j \geq L_j^*+1\right\}\right)\Bigg\}$$

$$+ \left(1-\lambda_i - b_i\mu_i\right) h_i(x_i, b_i), \qquad \forall \mathbf{x}, \mathbf{b}.$$

The constraint in the (**LP1**) is rather complex. In order to further simplify the constraint, the following steps are taken. First, the set $\mathcal{U}_i(\mathbf{x}, \mathbf{b})$ is replaced with

$$\mathcal{U}_i^{'}(x_i, b_i) = \left\{ \mathbf{a}_i = (a_i, t_i) \in \{0,1\}^2 \middle| a_i \leq \mathbb{I}\left\{b_i < B\right\}, \mathbb{I}\left\{x_i = K\right\} \leq a_i + t_i \leq 1 \right\}.$$

Second, by relaxing the constraint $d_j \leq x_j$ for all $j \neq i$, the set $\mathcal{D}_i(x)$ can be replaced with

$$\mathcal{D}_i^{'}(x_i) = \left\{ \mathbf{d}_i = (d_{i1}, \ldots, d_{in}) \in \{0,1\}^n \middle| d_{ii} \leq x_i, \sum_{j=1}^{n} d_{ij} \leq 1 \right\}.$$

Observe that $\mathcal{U}_i(\mathbf{x}, \mathbf{b}) \subseteq \mathcal{U}_i^{'}(x_i, b_i)$ and $\mathcal{D}_i(\mathbf{x}, \mathbf{b}) \subseteq \mathcal{D}_i^{'}(x_i)$. Similarly, the sets $\mathcal{U}_k(\mathbf{x}, \mathbf{b})$ and $\mathcal{D}_k(x)$ for $k \neq i$ are, respectively, replaced with

$$\mathcal{U}_k^{'}(x_i, b_i) = \left\{ \mathbf{a}_k = (a_k, t_k) \in \{0,1\}^2 \middle| a_k \leq \mathbb{I}\left\{b_i < B\right\}, \mathbb{I}\left\{x_i = K\right\} \leq a_k + t_k \leq 1 \right\},$$

and

$$\mathcal{D}_k^{'}(x_i) = \left\{ \mathbf{d}_k = (d_{k1}, \ldots, d_{kn}) \in \{0,1\}^n \middle| d_{ki} \leq x_i, \sum_{j=1}^{n} d_{kj} \leq 1 \right\}.$$

Note that $\mathcal{U}_k(\mathbf{x}, \mathbf{b}) \subseteq \mathcal{U}_k^{'}(x_i, b_i)$ and $\mathcal{D}_k(\mathbf{x}, \mathbf{b}) \subseteq \mathcal{D}_k^{'}(x_i)$. Hence, by taking this step, we effectively reduce the right hand side of the constraint in the (**LP1**). In the next step, the right hand side of the constraint is made dependent only on (x_i, b_i).

Let us define $\mathbf{x}_{-i} = \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$ and $\mathbf{b}_{-i} = \{b_1, \ldots, b_{i-1}, b_{i+1}, \ldots, b_n\}$. Now, in order to make the right hand side of the constraint independent of $\mathbf{x}_{-i}$ and $\mathbf{b}_{-i}$, we take the minimum over $\mathbf{x}_{-i}$ and $\mathbf{b}_{-i}$ for each given (x_i, b_i). Consequently, by using the new action space and simplifying, the constraint of the (**LP1**) will be a function of only x_i and b_i, and is written as:

$$
\begin{aligned}
h_i(x_i, b_i) + \rho \leq \pi_i x_i + \lambda_i \min_{\mathbf{a}_i \in \mathcal{U}_i'(x_i, b_i)} \Bigg\{ \kappa_i t_i + h_i(x_i + 1 - a_i - t_i, b_i + a_i) \Bigg\} + \\
(1 - \lambda_i - b_i \mu_i)\, h_i(x_i, b_i) + \min_{(\mathbf{x}_{-i}, \mathbf{b}_{-i}) \in \mathcal{B}(x_i, b_i)} \Bigg\{ \sum_{k \neq i} \pi_k x_k + \\
\sum_{k \neq i} \lambda_k \min_{\mathbf{a}_k \in \mathcal{U}_k'(x_i, b_i)} \Bigg\{ \kappa_k t_k + \pi_k w_k \mathbb{I}\left\{a_k + t_k = 0, x_k \geq L_k^*\right\} + \frac{\alpha}{\mu_k} \mathbb{I}\left\{a_k = 1, b_k \geq \tilde{b}_k^*\right\} \Bigg\} \\
+ b_i \mu_i \min_{\mathbf{d}_i \in \mathcal{D}_i'(x_i)} \Bigg\{ (1 - d_{ii}) h_i(x_i, b_i - 1) + d_{ii} h_i(x_i - 1, b_i) + \\
\sum_{j \neq i} d_{ij} \left(\frac{\alpha}{\mu_j} \mathbb{I}\left\{b_j \geq \tilde{b}_j^*\right\} - \pi_j w_j \mathbb{I}\left\{x_j \geq L_j^* + 1\right\} \right) \Bigg\} \\
+ \sum_{k \neq i} b_k \mu_k \min_{\mathbf{d}_k \in \mathcal{D}_k'(x_i)} \Bigg\{ (1 - d_{kk}) \left(-\frac{\alpha}{\mu_k} \mathbb{I}\left\{b_k \geq \tilde{b}_k^* + 1\right\} \right) + \\
d_{ik} \Big(h_i(x_i - 1, b_i + 1) - h_i(x_i, b_i) \Big) - d_{kk} \Big(\pi_k w_k \mathbb{I}\left\{x_k \geq L_k^* + 1\right\} \Big) + \\
\sum_{j \neq i,k} d_{kj} \left(\frac{\alpha}{\mu_j} \mathbb{I}\left\{b_j \geq \tilde{b}_j^*\right\} - \pi_j w_j \mathbb{I}\left\{x_j \geq L_j^* + 1\right\} \right) \Bigg\} \Bigg\}, \\
\forall x_i \leq K, b_i \leq B,
\end{aligned}
$$

where $\mathcal{B}(x_i, b_i) = \left\{ (\mathbf{x}_{-i}, \mathbf{b}_{-i}) \,\middle|\, \sum_{k \neq i} x_k \leq K - x_i, \sum_{k \neq i} b_k \leq B - b_i \right\}$.

The minimization over $\mathbf{x}_{-i}$ and $\mathbf{b}_{-i}$ can be taken out from above and be written as a separate mixed-integer program (**MIP**). We need to introduce binary variables

to replace the indicator variables as well as other binary and integer variables to remove the non-linear terms from the objective function. By taking all these steps, we will have an **MIP** with linear constraints and linear objective function, which is stated in the next section. The resulting **MIP** can be easily solved by CPLEX even with a huge number of variables and constraints. We denote this **MIP** program by $\mathbf{MIP}(x_i, b_i, h_i(x_i, b_i))$ to emphasize its dependency on x_i, b_i and $h_i(x_i, b_i)$. By plugging back the **MIP** into the (**LP1**), we have:

$$
\begin{aligned}
(\mathbf{LP2}) \quad & \max \quad \rho \\
& \text{s.t} \quad h_i(x_i, b_i) + \rho \leq \pi_i x_i + \lambda_i \min_{\mathbf{a}_i \in \mathcal{U}_i(x_i, b_i)} \left\{ \kappa_i t_i + h_i(x_i + 1 - a_i - t_i, b_i + a_i) \right\} \\
& \qquad\qquad + (1 - \lambda_i - b_i \mu_i)\, h_i(x_i, b_i) + \mathbf{MIP}(x_i, b_i, h_i(x_i, b_i)), \quad \forall x_i \leq K, b_i \leq B.
\end{aligned}
$$

Now we need to solve the (**LP2**) with ρ and $h_i(\cdot)$ as unknown variables. The structure of the (**LP2**) is equivalent to an average cost DP with state variables (x_i, b_i), and therefore is solvable by the relative value iteration algorithm. By implementing this decomposition scheme, we are approximating the (**DP**) which has $2n$ state variables by n separate smaller DP with only 2 state variables.

The optimal average cost obtained from value iteration algorithm is denoted by ρ_i^*. After implementing this algorithm, we also get $h_i(x_i, b_i)$ for all i, x_i and b_i. In the process of deriving the (**LP2**), some of the constraints in action space that exist in the original (**LP**) are relaxed. So the optimal average cost from the (**LP2**) should be a lower bound for the optimal average cost. This result is summarized in the following proposition.

Proposition 3.6.2 *The optimal objective function of the (**LP2**) gives a lower bound on the optimal average cost in the (**DP**); i.e., $\rho_i^* \leq \rho^*$ for each i. Consequently,* $\max_i \rho_i^* \leq \rho^*$.

3.6.3 The Mixed-Integer Program

To write the minimization over $\mathbf{x}_{-i}$ and $\mathbf{b}_{-i}$ as a separate MIP, we need to define binary variables to replace the indicator variables in the minimization. The first set of variables is:

$$z_k = \mathbb{I}\{x_k \geq L_k^*\} \text{ and } z_k' = \mathbb{I}\{x_k \geq L_k^* + 1\}, \qquad \forall k \neq i$$
$$r_k = \mathbb{I}\left\{b_k \geq \tilde{b}_k^*\right\} \text{ and } r_k' = \mathbb{I}\left\{b_k \geq \tilde{b}_k^* + 1\right\}, \qquad \forall k \neq i,$$

along with the following constraints:

$$\begin{aligned}
&x_k \geq z_k L_k^*, && x_k \leq (1 - z_k)(L_k^* - 1) + z_k M, && \forall k \neq i,\\
&x_k \geq z_k'(L_k^* + 1), && x_k \leq (1 - z_k')L_k^* + z_k' M, && \forall k \neq i,\\
&b_k \geq r_k \tilde{b}_k^*, && b_k \leq (1 - r_k)(\tilde{b}_k^* - 1) + r_k M, && \forall k \neq i,\\
&b_k \geq r_k'(\tilde{b}_k^* + 1), && b_k \leq (1 - r_k')\tilde{b}_k^* + r_k' M, && \forall k \neq i.
\end{aligned}$$

Note that M is a positive large number. These constraints assure that the variable takes the right value as the associated indicator variable does. The second set of binary variables is defined to remove the non-linear terms in the constraints:

$$\begin{aligned}
&f_k = z_k(1 - a_k - t_k), && \forall k \neq i,\\
&f_k' = r_k a_k, && \forall k \neq i,
\end{aligned}$$

along with below constraints:

$$
\begin{aligned}
&2f_k \le z_k + (1 - a_k - t_k) \le f_k + 1, && \forall k \neq i,\\
&2f'_k \le r_k + a_k \le f'_k + 1, && \forall k \neq i.
\end{aligned}
$$

Therefore, for given i, x_i, b_i and $h_i(x_i, b_i)$, the minimization over $\mathbf{x}_{-i}$ and $\mathbf{b}_{-i}$ can be summarized as follows:

$$
\begin{aligned}
\min & \sum_{k\neq i} \pi_k x_k + \sum_{k\neq i} \lambda_k \left[\kappa_k t_k + \pi_k w_k f_k + \frac{\alpha}{\mu_k} f'_k\right]\\
&+ b_i \mu_i \left[(1 - d_{ii}) h_i(x_i, b_i - 1) + d_{ii} h_i(x_i - 1, b_i) + \sum_{j\neq i} d_{ij} \left(\frac{\alpha}{\mu_j} r_j - \pi_j w_j z'_j\right)\right]\\
&+ \sum_{k\neq i} b_k \mu_k \Bigg[(1 - d_{kk}) \left(-\frac{\alpha}{\mu_k} r'_k\right) + d_{ki} \Big(h_i(x_i - 1, b_i + 1) - h_i(x_i, b_i)\Big)\\
&+ d_{kk} \left(\pi_k w_k z'_k\right) + \sum_{j\neq i,k} d_{kj} \left(\frac{\alpha}{\mu_j} r_j - \pi_j w_j z'_j\right)\Bigg]
\end{aligned}
$$

s.t.:

$$
\begin{aligned}
&\sum_{k\neq i} x_k \le K - x_i, \quad \sum_{k\neq i} b_k \le B - b_i,\\
&a_k \le B - b_i, \quad x_i - K + 1 \le a_k + t_k \le 1, && \forall k \neq i,\\
&\sum_{j=1}^{n} d_{kj} \le 1, \quad d_{ki} \le x_i, && \forall k,\\
&2f_k \le z_k + (1 - a_k - t_k) \le f_k + 1, && \forall k \neq i,\\
&2f'_k \le r_k + a_k \le f'_k + 1, && \forall k \neq i,\\
&x_k \ge z_k L^*_k, \quad x_k \le (1 - z_k)(L^*_k - 1) + z_k M, && \forall k \neq i,\\
&x_k \ge z'_k (L^*_k + 1), \quad x_k \le (1 - z'_k) L^*_k + z'_k M, && \forall k \neq i,
\end{aligned}
$$

$$b_k \geq r_k \tilde{b}_k^*, \quad b_k \leq (1 - r_k)(\tilde{b}_k^* - 1) + r_k M, \qquad \forall k \neq i,$$

$$b_k \geq r'_k(\tilde{b}_k^* + 1), \quad b_k \leq (1 - r'_k)\tilde{b}_k^* + r'_k M, \qquad \forall k \neq i,$$

$$x_k \text{ and } b_k \text{ are integer}, \qquad \forall k \neq i,$$

$$a_k, t_k, z_k, z'_k, r_k, r'_k, f_k, f'_k \text{ are binary}, \qquad \forall k \neq i,$$

$$d_{kj} \text{ is binary}, \qquad \forall k, j.$$

We still have some non-linear terms in the objective function of the **MIP**. Hence, the following binary variables are defined to transform it to a linear function:

$$s_{ij} = d_{ij} r_j, \qquad 2s_{ij} \leq d_{ij} + r_j \leq s_{ij} + 1, \qquad \forall j \neq i,$$

$$s'_{ij} = d_{ij} z'_j, \qquad 2s'_{ij} \leq d_{ij} + z'_j \leq s'_{ij} + 1, \qquad \forall j \neq i,$$

$$e_k = d_{kk} r'_k, \qquad 2e_k \leq d_{kk} + r'_k \leq e_k + 1, \qquad \forall k \neq i,$$

$$v_k = d_{kk} z'_k, \qquad 2v_k \leq d_{kk} + z'_k \leq v_k + 1, \qquad \forall k \neq i,$$

$$u_{kj} = d_{kj} r_j, \qquad 2u_{kj} \leq d_{kj} + r_j \leq u_{kj} + 1, \qquad \forall k \neq i, \quad j \neq i, k,$$

$$y_{kj} = d_{kj} z'_j, \qquad 2y_{kj} \leq d_{kj} + z'_j \leq y_{kj} + 1, \qquad \forall k \neq i, \quad j \neq i, k.$$

The last set of variables includes:

$$m_k = b_k r'_k, \qquad m_k \leq r'_k(B - b_i), m_k \leq b_k, (r'_k - 1)M + b_k \leq m_k, \qquad \forall k \neq i,$$

$$m'_k = b_k e_k, \qquad m'_k \leq e_k(B - b_i), m'_k \leq b_k, (e_k - 1)M + b_k \leq m'_k, \qquad \forall k \neq i,$$

$$n_{ki} = b_k d_{ki}, \quad n_{ki} \leq d_{ki}(B - b_i), n_{ki} \leq b_k, (d_{ki} - 1)M + b_k \leq n_{ki}, \qquad \forall k \neq i,$$

$$o_k = b_k v_k \qquad o_k, \leq v_k(B - b_i), o_k \leq b_k, (v_k - 1)M + b_k \leq o_k, \qquad \forall k \neq i,$$

$$p_{kj} = b_k u_{kj}, \quad p_{kj} \leq u_{kj}(B - b_i), p_{kj} \leq b_k, (u_{kj} - 1)M + b_k \leq p_{kj}, \qquad \forall k \neq i, \quad j \neq i, k,$$

$$q_{kj} = b_k y_{kj}, \quad q_{kj} \leq y_{kj}(B - b_i), q_{kj} \leq b_k, (q_{kj} - 1)M + b_k \leq q_{kj}, \qquad \forall k \neq i, \quad j \neq i, k.$$

By taking all these steps, we will have a mixed-integer program with linear constraints and linear objective function as follows:

$$\begin{aligned}
\min \sum_{k \neq i} \pi_k x_k + \sum_{k \neq i} \lambda_k \left[\kappa_k t_k + \pi_k w_k f_k + \frac{\alpha}{\mu_k} f'_k \right] \\
+ b_i \mu_i \Bigg[(1 - d_{ii}) h_i(x_i, b_i - 1) + d_{ii} h_i(x_i - 1, b_i) + \\
\sum_{j \neq i} \left(\frac{\alpha}{\mu_j} s_{ij} - \pi_j w_j s'_{ij} \right) \Bigg] \\
+ \sum_{k \neq i} \Bigg[\alpha(m'_k - m_k) + \mu_k n_{ki} \Big(h_i(x_i - 1, b_i + 1) - h_i(x_i, b_i) \Big) - \\
\mu_k o_k \pi_k w_k + \sum_{j \neq i,k} \mu_k \left(\frac{\alpha}{\mu_j} p_{kj} - \pi_j w_j q_{kj} \right) \Bigg]
\end{aligned}$$

s.t.:

$$\sum_{k \neq i} x_k \leq K - x_i, \quad \sum_{k \neq i} b_k \leq B - b_i,$$

$$a_k \leq B - b_i, \quad x_i - K + 1 \leq a_k + t_k \leq 1, \qquad \forall k \neq i,$$

$$a_k \leq B - b_i, \quad x_i - K + 1 \leq a_k + t_k \leq 1, \qquad \forall k \neq i,$$

$$\sum_{j=1}^{n} d_{kj} \leq 1, \quad d_{ki} \leq x_i, \qquad \forall k,$$

$$2f_k \leq z_k + (1 - a_k - t_k) \leq f_k + 1, \qquad \forall k \neq i,$$

$$2f'_k \leq r_k + a_k \leq f'_k + 1, \qquad \forall k \neq i,$$

$$x_k \geq z_k L^*_k, \quad x_k \leq (1 - z_k)(L^*_k - 1) + z_k M, \qquad \forall k \neq i,$$

$$x_k \geq z'_k(L^*_k + 1), \quad x_k \leq (1 - z'_k)L^*_k + z'_k M, \qquad \forall k \neq i,$$

$$b_k \geq r_k \tilde{b}^*_k, \quad b_k \leq (1 - r_k)(\tilde{b}^*_k - 1) + r_k M, \qquad \forall k \neq i,$$

$$b_k \geq r'_k(\tilde{b}^*_k + 1), \quad b_k \leq (1 - r'_k)\tilde{b}^*_k + r'_k M, \qquad \forall k \neq i,$$

$$2s_{ij} \leq d_{ij} + r_j \leq s_{ij} + 1, \qquad \forall j \neq i,$$

$$2s'_{ij} \leq d_{ij} + z'_j \leq s'_{ij} + 1, \qquad \forall j \neq i,$$

$$2e_k \leq d_{kk} + r'_k \leq e_k + 1, \qquad \forall k \neq i,$$

$$2v_k \leq d_{kk} + z'_k \leq v_k + 1, \qquad \forall k \neq i,$$

$$2u_{kj} \leq d_{kj} + r_j \leq u_{kj} + 1, \qquad \forall k \neq i, \forall j \neq i, k,$$

$$2y_{kj} \leq d_{kj} + z'_j \leq y_{kj} + 1, \qquad \forall k \neq i, \forall j \neq i, k,$$

$$m_k \leq r'_k(B - b_i), m_k \leq b_k, (r'_k - 1)M + b_k \leq m_k, \qquad \forall k \neq i,$$

$$m'_k \leq e_k(B - b_i), m'_k \leq b_k, (e_k - 1)M + b_k \leq m'_k, \qquad \forall k \neq i,$$

$$n_{ki} \leq d_{ki}(B - b_i), n_{ki} \leq b_k, (d_{ki} - 1)M + b_k \leq n_{ki}, \qquad \forall k \neq i,$$

$$o_k \leq v_k(B - b_i), o_k \leq b_k, (v_k - 1)M + b_k \leq o_k, \qquad \forall k \neq i,$$

$$p_{kj} \leq u_{kj}(B - b_i), p_{kj} \leq b_k, (u_{kj} - 1)M + b_k \leq p_{kj}, \qquad \forall k \neq i, \forall j \neq i, k,$$

$$q_{kj} \leq y_{kj}(B - b_i), q_{kj} \leq b_k, , (q_{kj} - 1)M + b_k \leq q_{kj}, \qquad \forall k \neq i, \forall j \neq i, k,$$

$$x_k, b_k, m_k, m'_k, e_k, v_k, o_k, , y_{kj}, p_{kj} \text{ and } q_{kj} \text{ are integer}, \qquad \forall k \neq i, \forall j \neq i, k,$$

$$a_k, t_k, z_k, z'_k, r_k, r'_k, f_k, f'_k, n_{ki} \text{ are binary}, \qquad \forall k \neq i,$$

$$d_{kj} \text{ is binary}, \qquad \forall k, j,$$

$$s_{ij}, s'_{ij} \text{ are binary}, \qquad \forall j \neq i.$$

3.6.4 Deriving Admission Policy from ADP

After obtaining $h_i(x_i, b_i)$ for each i, we can approximate the overall $h(\cdot)$ function according to:

$$h(\mathbf{x}, \mathbf{b}) \approx \sum_{i=1}^{n} h_i(x_i, b_i) \equiv \tilde{h}(\mathbf{x}, \mathbf{b}).$$

Once we know $\tilde{h}(\mathbf{x}, \mathbf{b})$, we can use the original (**DP**) to determine an action in each state $(\mathbf{x}, \mathbf{b})$. The rules that constitute the ADP policy are explained as follows:

The ADP Policy:

1. *In the case of arrival of a type-i patient, compare the costs associated with admission of the patient to the queue (if there is space in the waiting room), admission to the ward (if there is an empty bed), and transferring to another hospital, which are* $\tilde{h}(\mathbf{x}+\mathbf{e}_i, \mathbf{b})$, $\tilde{h}(\mathbf{x}, \mathbf{b}+\mathbf{e}_i)$, *and* $\kappa_i+\tilde{h}(\mathbf{x}, \mathbf{b})$, *respectively and choose the decision with the minimum cost.*

2. *In the case of discharge of a type-i patient, compare the costs associated with admission of the type-j patient from the queue (any type of which there is at least one patient waiting in the queue) and admitting no patient, which are* $\tilde{h}(\mathbf{x}-\mathbf{e}_j, \mathbf{b}-\mathbf{e}_i+\mathbf{e}_j)$; $\forall j : x_j \neq 0$ *and* $\tilde{h}(\mathbf{x}, \mathbf{b}-\mathbf{e}_i)$, *respectively and choose the decision with the minimum cost.*

3.7. Computational Experiments with Realistic Problem Instances

In this section, problem instances with four patient types are considered. Note that with four types of patients and a large number of beds, the optimal policy cannot be computed exactly due to the curse of dimensionality. In Section 3.7.1, two static admission policies, namely the *Bed Allocation* policy and the *Bid Price* policy which are based on the solution of the static model (**SM**) are described. Six instances of the problem are introduced in Section 3.7.2, while a comparative analysis

between the two static policies, i.e., the Bed Allocation and the Bid Price policies, the first-come-first-serve (FCFS) policy (as a benchmark), and the ADP policy over these six problem instances is reported in Section 3.7.3. Recognizing the difficulties associated with the implementation of the ADP policy in practice, Section 3.7.4 presents a priority cut-off policy that is inspired by the ADP policy (described in Section 3.6.4). In Section 3.7.5, we report on a second set of comparative analysis between the ADP policy, the ADP-based Priority Cut-off policy, and the current policy being used at the MNH. In Section 3.7.6, the robustness of the ADP policy performance respect to linear structure assumption for the waiting cost function is examined.

3.7.1 Two Static Admission Policies

Using the solution of the (**SM**), two heuristic admission policies are developed. The first heuristic policy uses $(\tilde{\lambda}_i^*, \tilde{b}_i^*)$ for all i. At any given time, the maximum number of beds occupied by type-i patients is $\tilde{b}_i^*$. Also, some of the new arrivals of type-i patients are transferred based on the adjusted arrival rate $(\tilde{\lambda}_i^*)$. This static policy is called the *Bed Allocation (BA)* policy and is summarized below.

The Bed Allocation (BA) Policy:

1. *Admit an arriving type-i patient if the number of occupied beds by type-i patients is less than* $\tilde{b}_i^*$.
2. *When all* $\tilde{b}_i^*$ *beds are occupied, and there is room available in the ED (i.e.,* $\sum_{i=1}^n x_i < K$*), admit the new arrival to the queue with probability of* $p_i = \frac{\tilde{\lambda}_i^*}{\lambda_i}$ *and transfer with probability of* $1 - p_i$*. If* $\sum_{i=1}^n x_i = K$*, we have no option except transferring the new arrival.*

An alternative policy is motivated by the revenue management literature, which is called the *Bid Price (BP)* policy. This involves using the dual variable of the first constraint in the (**SM**) (denoted by α). Recall that α represents the opportunity cost of occupying a bed per unit time. The average LOS for a patient of type i is μ_i^{-1} and hence, the average opportunity cost of admitting one type-i patient to a bed is $\alpha\mu_i^{-1}$. If the cost of transfer to another hospital is less than $\alpha\mu_i^{-1}$, the heuristic policy involves transferring all arrivals of type-i patients. This makes sense when there is no patient in the system ($\mathbf{x} = \mathbf{0}$) or when there is at least one available bed (note that these two are equivalent because there is no reservation in this type of policy). In the event that there are some patients present in the queue, however, a more precise policy would be to incorporate the patient's waiting cost. The average waiting cost of patients is approximated using average waiting time obtained from the (**SM**). From the solution of (**SM**), we know that, on average, type-i patients wait for $W_i^* = \frac{{L_i}^*}{\lambda_i^*(1-p_{k_i^*})}$ units of time. Hence, the waiting cost is estimated as $\pi_i {W_i}^*$. Using this average waiting cost, in the case that $\mathbf{x} \geq \mathbf{0}$, we let a patient from type i to enter system if $\alpha\mu_i^{-1} + \pi_i {W_i}^* \leq \kappa_i$ and transfer the new arrival, otherwise.

To complete the BP policy, a decision rule needs to be defined for admitting waiting patients in the queue when a bed becomes available. There are two possible options: using FCFS rule or prioritizing patients with higher waiting cost per period. To find the best policy, different combinations of FCFS and prioritization with and without incorporation of waiting costs have been tested. The priority rule incorporating waiting costs performed better than others in most of the numerical examples. Thus, our BP policy is summarized as follows.

The Bid Price (BP) Policy:

1. *If there is at least one bed available ($\sum_{i=1}^{n} b_i < B$), admit an arriving type-i patient to the ward if $\alpha\mu_i^{-1} \leq \kappa_i$ and transfer otherwise.*
2. *If there is no bed available ($\sum_{i=1}^{n} b_i = B$), admit a new arriving patient of type i to the queue if $\alpha\mu_i^{-1} + \pi_i W_i^* \leq \kappa_i$ and transfer otherwise.*
3. *If one bed becomes available, priority is given to the patients with highest waiting cost (as a tie-breaking rule, the patient with smaller index is admitted).*

3.7.2 Six Problem Instances

In light of the data summarized in Section 3.4, first we consider a *base case*, in which $\boldsymbol{\pi} = (70, 90, 145, 295)$, $\boldsymbol{\kappa} = 2\boldsymbol{\pi}$, and $B = 16$. Two more cases are developed through altering the service capacity by 25% in both directions, while the cost parameters remain the same. By doing so, we vary the level of congestion in the system to see its impact on the performance of the policy alternatives. The base case corresponds to case 2, whereas the problem instances with $B = 12$ and $B = 20$ correspond to case 1 and case 3, respectively. In cases 4-6, the waiting costs are increased for severe patients ($\boldsymbol{\pi} = (70, 90, 500, 600)$) as well as the patient transfer costs ($\boldsymbol{\kappa} = 3\boldsymbol{\pi}$) in order to observe how the admission policies respond to higher levels of patient sensitivity to ED boarding. For all six problem instances, the ED is assumed to accommodate a maximum of six boarding patients, i.e., $K = 6$.

The optimal policy for none of the six cases in our comparative studies can be found. Nevertheless, it is possible to compare the ADP policy to the other heuristic policies. To the purpose of comparison, a simulation model has been developed to find the average cost associated with a specific policy. The length of simulation

horizon is considered to be 10,000 days with 1,000 days of warm-up period. Using the simulation model, the average waiting time of all patients and the average transfer rates for each policy alternative are also reported. The simple averages do not reflect the true performance of each policy since transferring or ED boarding a mild patient is not as undesirable as transferring one severe patient. Therefore, the unit time waiting costs (π_i) are used as weights to compute the weighted averages.

3.7.3 Comparative Analysis I

Let us now turn to a performance comparison among the First-come-first-serve (FCFS), Bed Allocation (BA), Bid Price (BP), and ADP policies. The average total costs of the four admission policies for the six cases are depicted in Figure 3–8. In this figure, each plot corresponds to a case, and is located according to its congestion level (across the horizontal axis) and patient sensitivity to waiting (across the vertical axis). Evidently, the ADP policy produces the lowest average total cost in all cases. The other policy options fail to maintain low average total costs under all six patient sensitivity and congestion scenarios, e.g., the BP policy under case 3.

The average waiting times and the average rates of patient transfer associated with the admission policy options are depicted in Figure 3–9. The trade-off among these two performance measures is quite evident from this figure. The more a policy recommends transferring the patients to another hospital, the less the average waiting time experienced by the remaining patients. Note that the ADP policy seems to result in a more acceptable overall performance by balancing these two metrics. Even though the ADP policy does not produce the lowest average waiting time in all cases, its transfer rate is consistently reasonable.

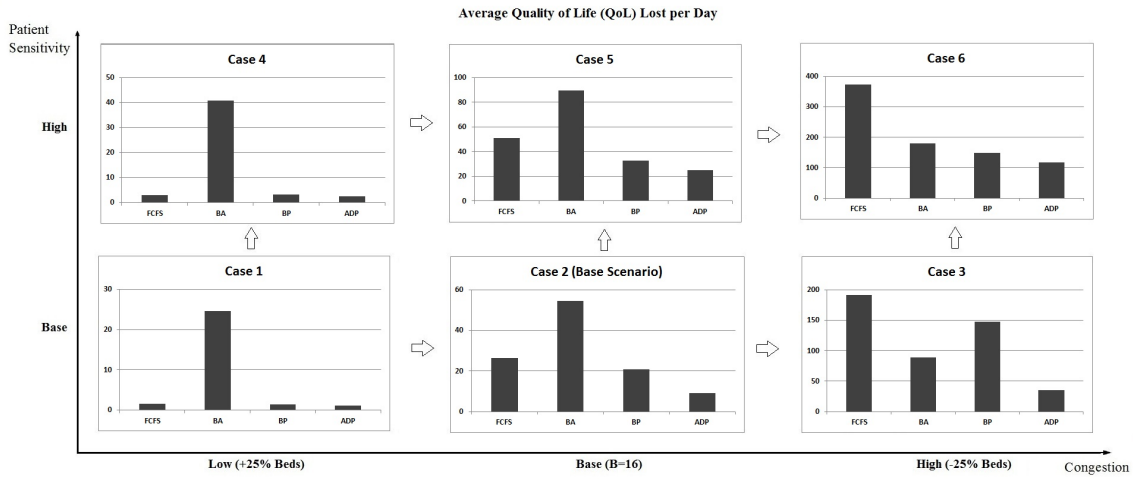

Figure 3–8: Average daily lost QoL – ADP policy versus static admission policies

In order to better display the comparison of the FCFS, BA, BP, and ADP policies, the plots in each of these two figures (and the two following figures) are not of the same vertical scale. Consequently, these figures do not highlight the true impact of increased congestion and patient sensitivity levels on the three performance measures.

By analyzing these figures, the following observations are made:

1. When the transfer cost increases (i.e., moving up in Figure 3–9), all policies – except BP – decrease the rate of transfers, which results in longer waiting times.
2. When the system is more congested (i.e., moving right in Figure 3–9), the transfer rates increase in all policies in order to avoid much longer ED boarding times.

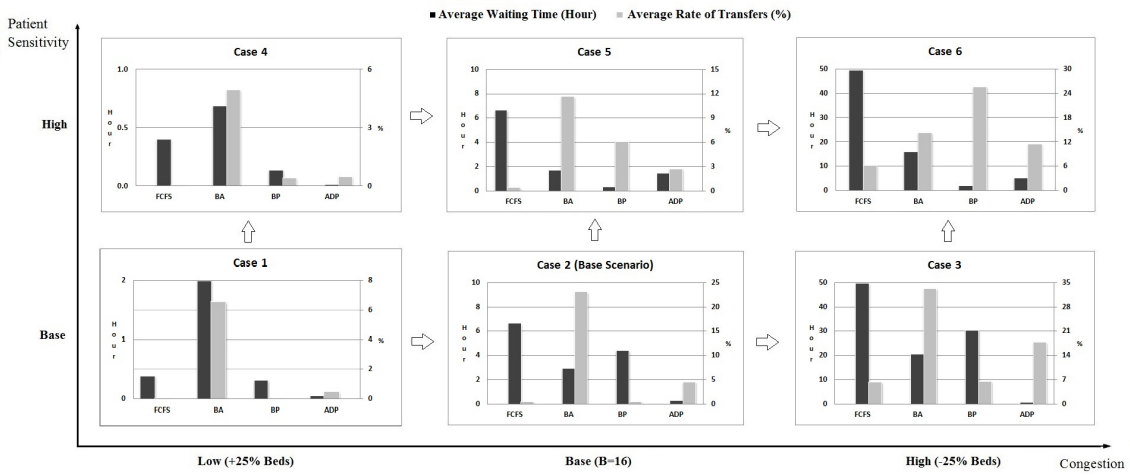

Figure 3–9: Average waiting time and rate of transfers – ADP policy versus static admission policies

3. The BP policy in cases 1-3 is reduced to a simple priority queue with no transfers. This happens due to the small value of α and average waiting times obtained from the (**SM**). In all these cases, the BP policy also dominates the FCFS policy.
4. The BA policy does not seem to be very promising. The total average cost of this policy is almost the highest in all cases, except in case 3 where its transfer rate is not acceptable.

The overall managerial insight from Figures 3–8 and 3–9 is that, as the congestion and patient sensitivity levels increase, the ADP policy performs increasingly better than the other policies in terms of achieving both lower costs and acceptable trade-offs between waiting times and patient transfers.

3.7.4 An ADP-based Priority Cut-off Policy

The ADP policy can be challenging to implement as it provides an action for every state of the system. Through a detailed analysis of the results of the ADP policy, however, we observe that often only a few of the system states are critical in nature. For instance, when there is only one bed available in the ward and a new patient arrives, the type of action we must take in response to the new arrival is crucial. Should we admit this new patient to the bed or save the last bed for arrival of a more severe patient in the future? As a more general question, how many beds should we reserve for severe patients by not admitting the mild patients? Or, is reservation necessary at all? In contrast, making the best decision when half of the service capacity is available seems to be trivial. A dynamic heuristic policy is developed in this section by following the ADP policy in the critical states and applying a simple policy such as FCFS rule in other states, which would be much easier to implement.

In order to ease exploring the structure of the ADP policy, the patients are first organized into two groups regardless of their disease; mild and severe patients. The patients in the mild group have lower waiting cost and shorter average LOS; while in the severe group patients are highly sensitive to waiting and they occupy the bed for longer time periods. It is crucial that we are able to cluster the patients into two distinct groups to develop this heuristic policy.

It is evident from the results of the ADP policy in six cases that the severe patients should be prioritized over mild patients. The ADP policy always admits a severe patient if there is an available bed. However, this is not true for the mild

patients. The ADP policy tends to reserve some beds for severe patients (by not admitting the mild patients to those bed) unless there is a high chance of a patient discharge in the near future. The chance of a discharge in the future depends on the patient mix in the ward, particularly the number of beds occupied by the severe patients.

Denote the aggregate number of the severe patients staying in the ward by b_{s}. The chance of a discharge is deemed high if $b_{\mathrm{s}} \leq \theta_1 B$, medium if $\theta_1 B < b_{\mathrm{s}} \leq \theta_2 B$, and low otherwise; $0 \leq \theta_1 < \theta_2 \leq 1$. In transferring the severe patients, there is a threshold on the cost associated with it that affects the transfer decision. The transfer cost in this heuristic policy is defined to be small if $\boldsymbol{\kappa} \leq \omega \boldsymbol{\pi}$ and to be large, otherwise. The values for these thresholds can be derived based on the ADP policy recommendations at the critical states of the system.

Note that some simplifications are required to obtain the thresholds from the ADP policy. For example, in developing this heuristic policy the number of patients in the queue is not incorporated in our admission decisions. This is justified by the results we obtained from the ADP policy in all six cases and it is mostly due to the low arrival rates of patients to the system in our examples. It is presumed that the queues are empty when a new patients arrives and consequently the decision is based only on the state of the ward. Therefore, the rules in this heuristic policy comply with the results of our illustrative example in Figure 3–4 of Section 3.5.

This heuristic policy is called the *ADP-based Priority Cut-off (PC)* policy because (i) it gives priority to certain types of patients, (ii) it changes behavior when

the state of the system surpasses the cut-off points. Priority cut-off policies are commonly used in the context of patient scheduling and healthcare capacity allocation (see for example; Esogbue and Singh (1976), Green et al. (2006), Ayvaz and Huh (2010), Mandelbaum et al. (2012b)). However, finding the best value of cut-off points (or thresholds) for this type of policy can be challenging. For the patient admission problem considered in this Chapter, the ADP policy could be used to find the structure of the PC policy as well as the appropriate threshold values. A general form of such ADP-based PC policy is stated below. Note that in the following, S denotes the number of beds reserved for severe patients.

The ADP-based Priority Cut-off (PC) Policy:

1. *When a severe patient arrives:*
   (a) *If at least one bed is available, admit the patient to the ward.*
   (b) *If all beds are occupied:*
      i. *if the transfer cost is small, then transfer the patient.*
      ii. *otherwise, admit the patient to the queue if the chance of a discharge is high and transfer the patient otherwise.*
2. *When a mild patient arrives:*
   (a) *If more than S beds are available, admit the patient to the ward (i.e., FCFS policy).*
   (b) *If between one and S beds are available:*
      i. *admit the patient to the ward if the chance of a discharge is high,*
      ii. *admit the patient to the queue if the chance of a discharge is medium,*
      iii. *transfer the patient if the chance of a discharge is low.*
   (c) *If all beds are occupied, admit the patient to the queue if the chance of a discharge is high and transfer the patient otherwise.*
3. *If a discharge occurs, the priority of admitting a patient to the ward is always given to the severe patients. If no severe patient is waiting in the queue, admission of a mild patient follows item 2.a.*

3.7.5 Comparative Analysis II

The second part of our analysis in this section involves comparing the ADP-Based Priority Cut-off (PC) policy and the current policy being used at the MNH with the ADP policy. The MNH policy has been briefly discussed in Section 3.1. It allocates a fixed number of beds to each patient type regardless of their level of severity and leaves some beds flexible to be used by all patient types. Let us denote the number of beds dedicated to stroke patient beds by b_{stroke}, number of beds dedicated to non-stroke patient beds by $b_{\text{non-stroke}}$, and the number of flexible beds by b_{flexible}. The patients are admitted to the beds until all the dedicated beds to their type and flexible beds are full. Then, they wait in the queue for a bed in the ward until the waiting time exceeds a threshold (denoted by T) in which case they have to be transferred. The hospital uses the same time threshold for all patient transfers. This policy, which is a static bed allocation policy, is summarized below.

The Current (MNH) Policy:

1. *When a patient arrives, admit the patient to the bed if any of the dedicated beds to that patient type is empty. If all the dedicated beds are full, the next option will be the flexible beds. If all the dedicated and flexible beds are occupied, then the patient waits in the queue.*
2. *If the wait time for a patient in the queue exceeds T, the patient is transferred to another hospital.*

In the experiments, the current policy uses the following parameters:

- In cases 1 and 4, we have $(b_{\text{stroke}}, b_{\text{non-stroke}}, b_{\text{flexible}}) = (8, 8, 4)$.
- In cases 2 and 5, we have $(b_{\text{stroke}}, b_{\text{non-stroke}}, b_{\text{flexible}}) = (6, 6, 4)$.
- In cases 3 and 6, we have $(b_{\text{stroke}}, b_{\text{non-stroke}}, b_{\text{flexible}}) = (5, 5, 2)$.

- In cases 1-3, we have $T = 48$ hr, and in cases 4-6, we have $T = 72$ hr.

Also, note that the thresholds of the ADP-based PC policy explained in Section 3.7.4 vary with the cost parameters. By examining the results of the ADP policy for the six cases, the following parameters are observed:

- In all cases, the threshold associated with the transfer cost is $\omega = 2$.
- In cases 1-3, we have $S = 1$, $\theta_1 = 1/4$, and $\theta_2 = 1/2$.
- In cases 4-6, we have $S = 4$, $\theta_1 = 1/2$, and $\theta_2 = 3/4$.

Since the waiting costs of the severe patients are much higher in cases 4-6, the number of beds reserved for them is larger. Also, the larger transfer costs in cases 4-6 lead to higher thresholds for evaluating the likelihood of having an available bed in the future.

The average total costs of the ADP, PC and MNH policies are depicted in Figure 3–10. The ADP policy has the lowest average cost in all cases, whereas the costs associated with the PC policy are consistently within an acceptable range of the ADP policy. The difference between the ADP policy (or PC policy) and the MNH policy is more pronounced when the patients are more sensitive to waiting and service capacity is limited (i.e., cases 2-3 and 5-6).

The average waiting time and average rates of patient transfer for the three policies are shown in Figure 3–11. The PC policy is more conservative than the ADP policy in terms of patient transfers. In all cases, it transfers fewer patients and consequently it has higher average waiting times. Compared to the current policy, the ADP policy decreases the waiting time significantly while its transfer rates are slightly higher in some cases. The PC policy, however, reduces the wait times in most

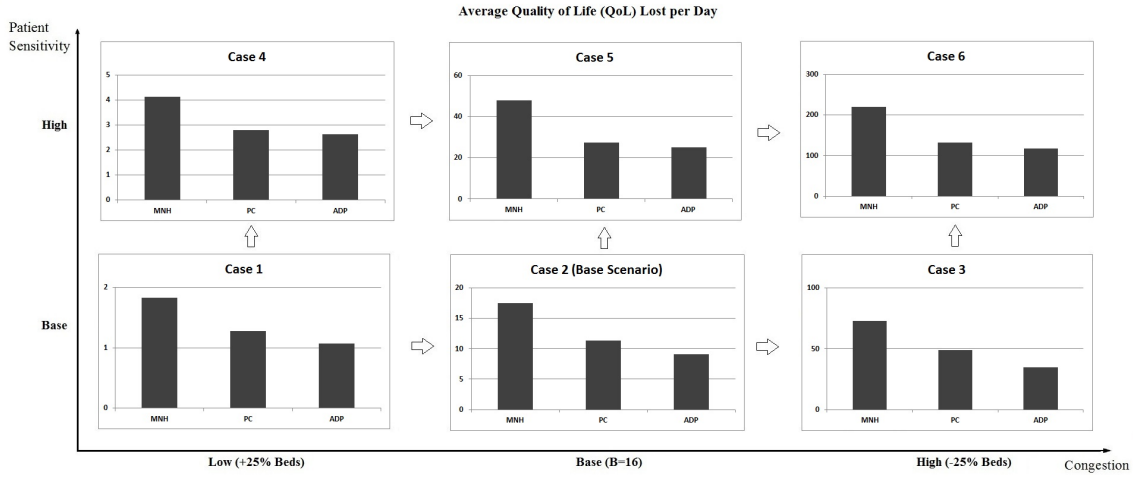

Figure 3–10: Average daily lost QoL – ADP policy versus practical admission policies

cases by transferring the same or less number of patients. Hence, it could be utilized as an efficient and practical policy by the hospital to improve the performance of the ward in terms patients' health outcomes. It is also important to note that the PC policy generates the second lowest average costs over six cases compared to the static policy alternatives (i.e., BA, BP, and FCFS policies) in Section 3.7.3.

3.7.6 Non-linear Cost Functions

Since the patient's health status may deteriorate faster in some cases as the waiting time increases, non-linear waiting cost functions are considered in this section. To this end, a piecewise-linear increasing convex function for the waiting costs of patients is assumed. The time is divided into three-hour intervals and in each interval the waiting cost is a linear function of time with a slope that is increasing from one interval to the next one. However, we need to linearize this cost function so as to implement the admission policies. Hence, a Regression through Origin (RTO)

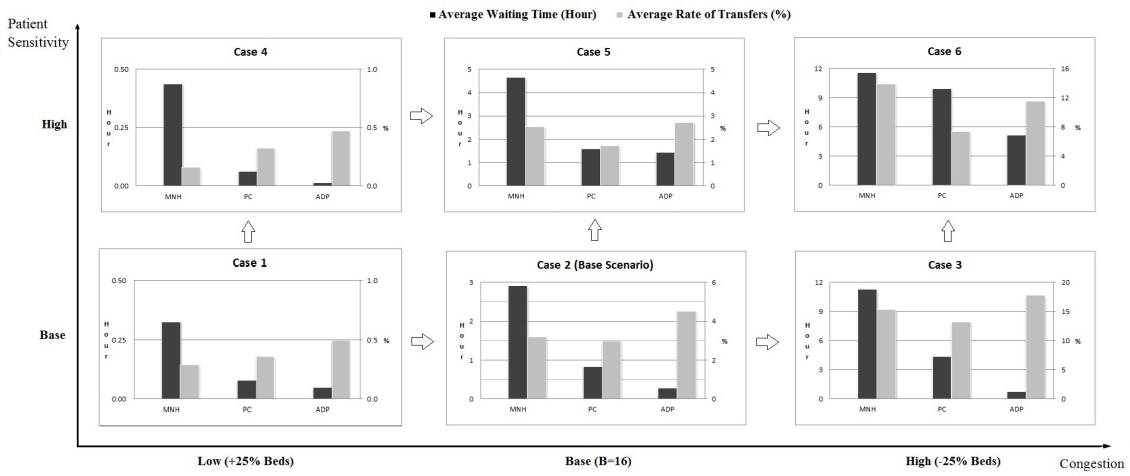

Figure 3–11: Average waiting time and rate of transfers – ADP policy versus practical admission policies

model is developed to find the best linear function that fits to the data points obtained from each non-linear cost function. In order to compare the performance of the admission policies under linear and non-linear waiting costs, the parameters of non-linear function are chosen such that the slope of fitted linear function is equal to the waiting cost per unit time (π_i) in those cases that were considered in Section 3.7.3. The simulation model is then used to calculate the average total cost in both scenarios.

The results are shown in Table 3–5. In this table, the percent increase in the total cost associated with each policy, when the waiting costs are incurred according to a non-linear function is reported. The last two columns of this table show the percent improvement achieved by the ADP policy in each scenario over the best of the other policy options. A negative percentage implies that the ADP policy is dominated by another heuristic policy. From the table, it can be concluded that

the performance of the ADP policy remains robust to the change in the structure of waiting costs in almost all cases. Except in case 4, in which there is enough service capacity, the percent improvement of the ADP policy over other policies has in fact increased.

Case	Cost Increase (%)					ADP Improvement (%)	
	FCFS	BA	BP	PC	ADP	Linear	Non-linear
1	63	73	63	41	29	27	42
2	98	53	94	56	33	57	70
3	134	77	126	78	35	61	71
4	64	102	76	96	91	12	-3
5	99	102	92	114	73	24	31
6	137	131	98	151	87	22	26

Table 3–5: Robustness of the ADP policy respect to non-linearity of waiting cost

3.8. Conclusion

In this Chapter, an admission control and bed allocation problem that incorporates the differentiating features of neurology wards has been considered. From a modeling perspective, an average cost DP that assumes none of the beds in the ward are earmarked to certain patient types was presented. It was shown that the optimal policy for admitting patients from the ED is dynamic and it depends on the state of the system. To overcome the curse of dimensionality, an ADP was proposed that uses some information from a static model, which is developed utilizing queuing theory principles. To the best of the author's knowledge, the ADP for the average

cost problem has not been fully explored theoretically. Some examples include Roubos and Bhulai (2010) and Roubos and Bhulai (2012) that use ADP in controlling queues with application to call centers.

The numerical results on realistic-size problem instances, based on Montreal Neurological Hospital, revealed that the admission policy suggested by the ADP works very well compared with the other heuristic policies. Recognizing the managerial challenges in implementing the fully state dependent ADP policy, an ADP-based priority cut-off policy was developed that performs quite well. It must be emphasized that the structure of this heuristic policy is highly dependent on the results of the experiments for the six problem instances that were considered in the comparative analysis.

The current admission policy at the hospital involves dedicating six beds to stroke patients and six beds to non-stroke patients, while leaving four beds flexible for both patient types. Furthermore, a patient transfer request is triggered after 48 hours of ED boarding. In contrast, the proposed ADP policy does not use earmarked beds and decides to transfer the patient at the time of arrival, considering the state of the system. By comparing these two policies, it is shown that the current policy can be 70-110% worse than the ADP policy in terms of average HRQoL lost per day. Also, the ADP policy can decrease the average boarding time in the ED (especially when there are limited number of beds available such as in case 3 and case 6 of our comparative analysis) significantly without affecting the average rate of patient transfers. Thus, the following insights are provided for neurology ward managers: (i) it is better to decide whether or not to transfer a patient to another hospital

immediately upon arrival and by taking into account the state of the system, (ii) dedicating neurology ward beds to patient types can worsen average ED boarding times, (iii) if the managers prefer to use an earmarking strategy, it is recommended to do so based on the level of severity of the patients condition rather than their disease (i.e., along the lines of the PC policy).

One of the limitations of the modeling framework of this study and the solution approach is the stationary arrival process assumption. The stationary case was chosen to simplify the analysis and the exposition of the material, as customary in the healthcare operations literature (Patrick et al., 2008). However, this study can be adapted to deal with non-stationary arrival processes. A well-studied technique to deal with non-stationary arrival in queueing control is point-wise stationary approximation (PSA). The PSA approach uses solutions from stationary systems as building blocks for non-stationary systems (Green and Kolesar (1991) and Yoon and Lewis (2004)). A PSA solution can be easily constructed based on the static queueing approximation and the DP decomposition introduced in our paper.

The modeling framework proposed in this chapter is based on two more simplifications. First, a small percentage of the patients with neurological conditions can be admitted directly to the ward for elective surgeries, while this study was confined to the patients who are admitted through the ED. Second, some patients, e.g., severe stroke patients, require intensive care for stabilization prior to being admitted to the ward, which is missed in the model. Their LOS in the neuro-ICU, however, is most often 48 hours with fairly low variability. Extensions to the model to relax these two assumptions constitute fruitful avenues for future research.

In closing, the problem that was studied shares similarities with general multi-class queuing problems, and hence this research can have potential implications in a more general domain than the healthcare context that constitutes the focus of this chapter.

CHAPTER 4
The Specialization of Healthcare Services

4.1. Introduction

Networks of multiple hospitals are becoming very popular structure in the healthcare sector (Yonek et al., 2010). Hospitals seek to improve efficiencies, remain competitive and increase the chance of survival through mergers and consolidation. More organizational changes are expected to take place as a consequence of rising hospitals costs and healthcare budget cuts (`www.ft.com`). In general, there is a common belief that mergers can enhance operational efficiency and produce economic benefits, but the outcome is not always guaranteed. Two examples of successful mergers in Canada are The Ottawa Hospital and Trillium Health Partners. Significant operational improvements have been reported in these hospitals after the mergers. The Trillium Health Partners has experienced a seven percent reduction in waiting times in EDs over all its four sites (`www.hospitalnews.com`).

However, there have always been different opinions among both practitioners and researchers about the success of mergers and the difficulty of implementing changes at the merger sites. As merging and centralization of services are often observed in the healthcare sector, it does not always appear to be the best solution. The hospitals that provide specific types of care to patients are expected to benefit from economies of focus and improve the quality of the care provided by them. This is achieved by organizing the care around patients groups, such as the breast cancer clinics or

the clinical pathways for diabetes patients (Vanberkel et al., 2012). But as it has been realized from many studies this might not be ideal in every situation. Thus, a main challenge faced by policy makers in the process of designing the structure of multi-hospital networks is finding the best system configuration that improves the quality of care and efficiency of hospitals.

A network of hospitals is no different than a queuing network. Patients enter the network, request for medical services from – possibly multiple – servers, renege or wait for a server if that server is busy, and leave the system after completion of their treatment. Examining the queueing theory literature, we observe a significant number of studies that have focused on designing the queueing systems in terms of customer routing and server flexibility level. Many studies compare possible scenarios from dedicated (or specialized, or decentralized) to fully flexible (or pooled, or centralized, or diversified) configurations. The dedicated system refers to an independent single queue while the pooled system operates as a multi-server queueing system with a single queue. It has been shown that when service and demand distributions are homogenous, the pooled system always dominates the dedicated system (Smith and Whitt (1981), Benjaafar (1995), Joustra et al. (2010), Ata and Van Mieghem (2009)). However, if the system is not homogenous, the pooled system is not always preferred (Buzacott (1996), Smith and Whitt (1981), Mandelbaum and Reiman (1998), Dijk and Sluis (2008), and Van Dijk and van der Sluis (2009)).

In a healthcare application to health care setting, Vanberkel et al. (2012) examines the impact of pooling on the efficiency of hospitals. The trade-off between economies of scale resulting from centralization versus economies of focus resulting

from decentralization is studied in this paper. The centralization of healthcare service in this study implies developing healthcare centers that serve all types of patient and decentralization refers to centers that offer a limited range of services. They have found that the decision of dividing a centralized center requires careful consideration of center load, patient mix, and variability of service times.

Tiwari and Heese (2009) studies a network of two hospitals with the presence of one competitor and answers the question of when specialization of service is preferred and when the network of hospitals is better off by remaining diversified from a profit maximizing perspective. Mahar et al. (2011) studies the problem of locating a specialized healthcare delivery systems, while taking into account both financial and patient service level aspects. For further literature on this topic, the reader is referred to Section 2.1.

Reviewing the related literature reveals that one right answer to the question of pooling versus unpooling, specialization versus diversification, or dedication versus flexibility does not exist. This study aims at shedding additional lights to the answer of this question in the context of healthcare services. It provides extra information pertinent to the situations when specialization of service could be beneficial.

Before elaborating on the motivation of this study, its differentiating features are highlighted. First, the possible network designs of this study are different from the ones considered in prior studies. While the specialization scenario of this study is the same, the alternate scenario (i.e., diversification) is quite different. In our diversification scenario the hospitals work completely independent of each other and unlike the pooled scenario every hospital has its own resources and patient inflows.

Due to this specific feature the comparison of each scenario requires finding the best allocation of resources between the hospitals. This implies that in the inner layer of each comparison, there is an optimization problem for which an efficient solution methodology is developed.

Second, this study considers a wide range of characteristics belonging to a multi-hospital network and introduces a new set of parameters for describing the properties of a network. Third, it considers the impact of blocking in the process of providing service to patients, which is very relevant and is sometimes missing in healthcare research. In this study, the refusal rate of patients to the network is considered as the performance measurement of scenarios while other studies use average waiting time of patients or throughput rate of patients in the system.

Fourth, this study examines the simultaneous improvement of all hospitals involved in a restructuring process that, to the knowledge of the author, has not been considered in earlier papers. Besides searching for the situations where specialization could improve the network-level performance, this study identifies the conditions where specialization helps all the hospitals of the network increase their operational efficiency.

4.2. Motivation

As a pending restructuring of two existing sites of McGill University Health Center (MUHC), the Montreal Neurological Hospital (MNH) and the Montreal General Hospital (MGH) plan to collaboratively provide care to stroke patients. In the proposed configuration, each of these two hospitals will offer certain level of stroke care; secondary stroke care (at MGH) and tertiary stroke care (at MNH). A patient

that is diagnosed as a stroke case will be taken to the ED of one of these hospitals. Upon arrival at the ED, the patient is triaged by a special team of healthcare professionals to determine the level of care needed by the patient. The level of care decides to which hospital's ward the patient will ultimately be admitted. But in the current configuration, both hospitals are capable of providing both levels of care and stroke patients are admitted to the ward of the hospital they arrive at.

4.2.1 Stroke Patients Flow

In general, stroke patients are categorized into two groups: (1) hemorrhagic stroke patients; and (2) ischemic stroke patients. Hemorrhagic stroke patients normally need neuro-surgery or other types of neuro-intervention, which is considered as tertiary care, and thus, according to the proposed configuration, the patient should be treated at MNH. In the case of an ischemic stroke, if the time passed from the stroke is less than a certain amount of time (three hours in the current medical protocol), the patient is eligible for receiving tPA, which will be provided only at MNH. A small percentage of ischemic patients are tPA eligible. There is also a very small percentage of ischemic patients that are not tPA eligible, but may need an intervention, and as mentioned earlier, is considered as tertiary care provided only at MNH. All other ischemic patients require secondary care, which be provided at MGH when the proposed reconfiguration is implemented.

All the stroke patients, after receiving neuro-intervention or tPA, are transferred to the ICU and will be monitored carefully for a certain amount of time (i.e. 24 hours) to assure that the patient's health status is stabilized. After the stabilization, the patient is admitted to the stroke ward to continue receiving the acute care. Once the

patients finish their stay at the stroke ward, they are transferred to the Alternate Level of Care Unit (ALCU), where they wait until they are discharged to home or until a bed in rehabilitation center or long-term care facility becomes available for them. The care the patients receive in the ALCU is less intensive compared with the care they receive in the stroke ward and the ratio of healthcare professionals to patient in this unit is significantly lower.

The proposed configuration is an example of specialization of healthcare services. Each hospital will focus on providing a specific level of care in this scenario. The MNH will serve only severe stroke patients who need tertiary level of care. The MGH will admit only mild stroke patients who require secondary level of care. The flow of patients in this specialized setting is shown in Figure 4–1.

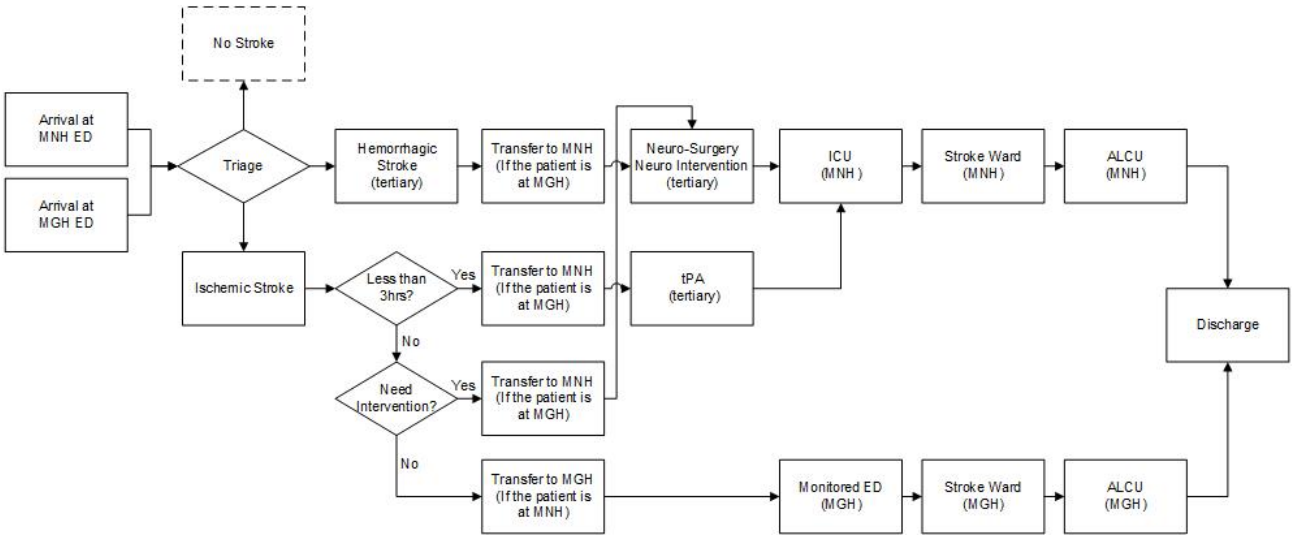

Figure 4–1: Patient flow in the specialized network scenario

In the current scenario, which is called diversified scenario, both hospitals receive both types of patients. The flow of the patients inside the hospitals is the same. However, no hospital is dedicated to serve a specific type of stroke patients. In Section 4.2.2, the two configuration scenarios are evaluated using a simulation model.

4.2.2 Simulation Study of Network Configuration Scenarios

Based on the stroke patients flow in the two hospitals, which has been described in Section 4.2.1, a discrete-event simulation model has been developed. The data of the mild and severe stroke patients is used for the arrival rate and the LOS of secondary and tertiary patients receptively, which can be found in Tables 3–1 and 3–2. Since the related data for the MGH is not available, for the sake of analysis, we use the same arrival rate and length of stay of the patients that arrive at MNH. Table 4–1 shows the number of beds that are currently available at different units of each hospital.

Table 4–1: Current number of beds at the MNH and MGH

Unit	Number of beds
MNH Neuro-ICU	3
MGH Monitored ED	3
MNH Stroke Ward	12
MGH Stroke Ward	8
MNH ALCU	6
MGH ALCU	6

One of the key performance measurements of the system is the average waiting time that patients experience before they are admitted to the stroke ward. As emphasized before, the treatment of stroke patients is very time-sensitive and it is highly desirable to keep the waiting time at any phase of their treatment as minimum

as possible. Therefore, using the simulation model, the average time a patient has to wait, on average, for a bed in the stroke ward are reported in Table 4–2.

Table 4–2: Average waiting time for a stroke ward bed (simulation results) – current bed allocation

Scenario	Average Waiting Time (hr)	
	MGH	MNH
Diversified	11.01	0.00
Specialized	15.17	0.01

The total number of stroke beds in two stroke wards is currently 20. The performance of the system can be improved by changing the allocation of beds between the two sites. The best bed allocation under each scenario can be obtained by exploring all the possible alternatives – the number of all allocation solutions is constrained by the total number of beds, i.e., 20. In searching for the optimal bed allocation, we minimize the maximum average waiting times of the two sites. By doing so, we decrease the waiting time for both secondary and tertiary stroke patients simultaneously. The performance of the system under two scenarios when the beds are optimally allocated is reported in Table 4–3.

Table 4–3: Average waiting time for a stroke ward bed (simulation results) – optimal bed allocation

Scenario	Average Waiting Time (hr)	
	MGH	MNH
Diversified	0.37	0.56
Specialized	1.06	0.32

In the optimal solution, the average waiting time for the mild stroke patients at MGH decreases significantly. However, this happens at the expense of slight

increase in the waiting time of severe stroke patients at MNH. Comparing the results in Tables 4–2 and 4–3 shows that the diversified configuration of the hospitals is preferred even if the current bed allocation remains intact. By specialization of the two hospitals the maximum waiting times of patients at two sites would increase. Specially, the MGH site is worse off in both current and optimal bed allocation, while the MNH would benefit from it if the optimal bed allocation is deployed. In the current bed allocation, the MNH is indifferent between the specialization and diversification scenarios.

This conclusion might seem counter-intuitive. It is presumed that by reducing the variability in service times of a queue, the average waiting time decreases. By specialization of the two sites we practically assign each site to deal with only one type of patient. So the variance in the LOS as well as the average waiting time for all patients should be reduced. However, this contradicts the results from the simulation model. It sounds that there are some other parameters than service time distribution that are playing role in the determining the performance of a multi-hospital network. In the next sections, the problem of multi-hospital network design is analyzed through a comprehensive experimental study. But before that, the resource allocation problem in a multi-site network is formulated so that the performances of configuration scenarios are compared at their optimal levels.

4.3. Multi-Site Healthcare System

Consider a network of hospitals that provides medical care to certain types of patients. Each type of patients requires a specific type of care. The patients are categorized into distinct groups based on the type of care they need. Let us denote

the patient group (or the patient type) by index i. Each hospital in the network is capable of providing all types of care. The hospitals are indexed by j.

We focus on the process of providing service only to inpatients. Inpatients stay in hospital for longer periods of time compared with outpatients and coordination of their care process requires additional efforts by hospital staff. Inpatients normally visit more than one unit during their stay in hospital and they flow inside the hospital according to their clinical pathways. Once their service is complete at one unit, they move to the next unit in their clinical pathways unless there is no service capacity available at the destination unit. In this case, the patients are prevented from moving forward (i.e. blockage occurs) and they continue using the resources of the current stage.

Inpatients enter hospitals through the ED. When their service at the ED is finished a request is generated for their admission to other units of hospital. A simple inpatient's path in a hospital can be the following: transferring to the Intensive Care Unit (ICU), being admitted to the Acute Care Unit (ACU), transferring to the Alternative Level of Care Unit (ALCU) – which might also be called the Post-Acute Care Unit (PACU) – and discharging from the hospital.

We assume a tandem queueing network in this study with no buffers between the stages. This means that the order according to which patients visit the hospital units is the same for all patient types. However, patients might visit any stage of their clinical pathway more than one time due to medical complications. Such considerations are excluded from the scope of this study for the sake of simplicity. Figure 4–2 provides an overview of a patient's clinical pathway in a hospital of the

network. Let us denote the stages in a patient's clinical pathway by k. The average time that a type-i patient stays at stage k is $\frac{1}{\mu_{ik}}$. Note that the LOS of patients is independent of location of hospitalization.

Figure 4–2: Clinical path for a patient in the hospital

4.4. Multi-Site Bed Allocation Problem

The patients are assumed to arrive at the network according to a Poisson process with the rate of λ per unit time. The arrival of patients to each hospital is independent of other hospitals and is a function of network configuration. We consider two configurations for the flow of patients in the network. In the first configuration, which is called *Diversification*, all types of patient might arrive at any hospital of the network. The type-i patients arrive at hospital j with the rate λ_{ij} per unit time. Denote the matrix of arrival rates by Λ; $\Lambda_{i,j} = \lambda_{ij}$. The rate of total arrivals at hospital j is $\lambda_{.j} (= \sum_i \lambda_{ij})$.

In an alternative scenario, which is called *Specialization*, all the patients of the same type will be served in one hospital. This implies that each hospital in the network is dedicated to serve a specific type of patients (or more than one type if the number of patient types is more than the number of hospitals). All the patients from one type that arrive at different hospitals in *Diversification* scenario, are now redirected to the hospital that is assigned to them. The hospital managing all type-i patients will receive patients with the rate of $\lambda_{i.} (= \sum_j \lambda_{ij})$ and will be specialized in offering service to a specific type of patients.

There is a limited number of beds available at each stage of the clinical path to be distributed among the hospitals of the network. The beds available at each stage are exclusively used for that stage and cannot be relocated to other stages. Denote the total number of available beds at stage k by B_k and the beds dedicated to the stage k at hospital j by b_{jk} after the allocation decisions are made. Note that $\sum_j b_{jk} = B_k$. The objective of Multi-Site Bed Allocation (**MSBA**) problem is to find the right number of beds assigned to each stage for all hospitals such that the performance of the whole network is maximized.

The performance of the network is measured by the maximum blocking probability of all sites. The acute care received by the inpatients is a very specialized treatment and unavailability of resources to provide the required care is very costly from a patient health perspective. The blocking probability demonstrates the proportion of patients whose access to the system is denied because there is no resource available at the first stage. By minimizing the maximum blocking probability of all sites we try to find the bed allocation decisions that minimizes the refusal rates at all hospitals simultaneously. As an alternative measurement, the (weighted) average of blocking probabilities could be considered as well. However, the major drawback of this criterion is that the difference between blocking probability of hospitals could be inappropriately large. To avoid any inequity issues between the hospitals or the patients, the maximum of blocking probabilities is chosen as the objective function of the **MSBA** problem.

Every solution of the **MSBA** problem can be shown by matrix $\mathbf{B}$; $\mathbf{B}_{j,k} = b_{jk}$. The blocking probability of hospital j depends on the bed allocation solution (jth

row of matrix $\mathbf{B}$ denoted by $\mathbf{b}_j$) and the scenario according to which the structure of the network is configured. Let us denote the blocking probability of hospital j for a given $\mathbf{b}_j$ by $p_j^{(\cdot)}(\mathbf{b}_j)$, where the superscript $(\cdot)$ refers to the network design, i.e., D for *Diversification* scenario or S for *Specialization* scenario. The **MSBA** problem for scenario $(\cdot)$ is defined as follows:

$$
\begin{aligned}
(\mathbf{MSBA}) \quad \text{Minimize} \quad & \max_j \{p_j^{(\cdot)}(\mathbf{b}_j)\} \\
\text{Subject to} \quad & \sum_j b_{jk} = B_k, \forall k, \\
& b_{jk} \text{ integer}, \forall j, k.
\end{aligned}
$$

The solution of the **MSBA** problem is denoted by $\pi^{(\cdot)}$, i.e., $\pi^{(\cdot)} = \min_{\mathbf{B}} \max_j p_j^{(\cdot)}(\mathbf{b}_j)$ with respect to the bed allocation constraints.

4.5. Blocking Probability Estimation in a Queueing Network

The exact blocking probabilities of a tandem queue without buffer space between stages are obtained through solving the system of steady-state equations. Finding the exact solution of this system of equations becomes very complex and the computation time increases exponentially as the state of the system expands. Therefore, approximation techniques are used to estimate the blocking probability for these queueing networks.

By examining the literature, several papers are found that develop approximation schemes and heuristic algorithms for estimating the blocking probabilities in different types of queues. Most of these papers take advantage of limiting assumptions. Examples of papers that study single server queues include: Hunt (1956),

Hillier and Boling (1967), Takahashi et al. (1980), Suri and Diehl (1984), Perros and Altiok (1986), Altiok and Perros (1987), Gershwin (1987), Brandwajn and Jow (1988), Lee and Pollock (1990), Dallery and Frein (1993), Perros (1994), Lee et al. (1998), Abadi et al. (2000), and Balsamo et al. (2001). Multi-server queues are studied by Hershey et al. (1981), Weiss et al. (1982), Weiss and McClain (1987), El-Darzi et al. (1998), Koizumi et al. (2005), and Bretthauer et al. (2011).

Among the cited studies, the recent paper by Bretthauer et al. (2011) shares the most similarities with the settings of the problem considered in our study. They study applications of tandem queue networks with blocking in a healthcare environment and develop a new heuristic to estimate the blocking probabilities in all stages of the network. An extension of the algorithm that can be applied to queueing networks with general routing is also developed. They compare the solution of their proposed heuristic with the exact solution and that of other heuristics that exist in the literature. Through extensive computational efforts, they show that the average percent error in the estimated blocking probability of first stage in a tandem queueing network is very small when compared with other heuristics. As a matter of fact, the average percentage errors for two- and three-stage systems (where the exact solution can be determined) is less than 5%.

Due to the good performance of the proposed heuristic approach in Bretthauer et al. (2011), their algorithm is used in finding the blocking probabilities in our problem. Note that this thesis does not intend to contribute to the literature of this area and thus their algorithm is only adjusted according to the settings of the problem under study. A summary of their heuristic algorithm follows.

4.5.1 Heuristic By Bretthauer et al. (2011)

The heuristic algorithms used to find the blocking probabilities in queuing networks often attempt to assess the effect of blocking in different stages of a network on the blocking probability of the first stage. Bretthauer et al. (2011) use two ideas: (i) adjusting the number of servers; and (ii) adjusting the service rates at each stage; to incorporate the fact that some servers are occupied by the patients whose service is complete but are unable to move forward because there is no free bed at the next stage.

To explain their proposed algorithm in more details, the same notations introduced by the authors will be used in this section. Consider a tandem queue with n stages. The number of servers at stage k is s_k and the service rate per server is μ_k. If we isolate stage k from other stages, the type of queue is $M/G/s_k/s_k$. Given the arrival rate λ, the probability of the system being full in such a queue is given by (Gross et al., 2008):

$$\pi(\lambda, \mu_k, s_k) = \frac{(\lambda/\mu_k)^{s_k}/s_k!}{\sum_{j=0}^{s_k}(\lambda/\mu_k)^j/j!} \tag{4.1}$$

The flow rates between stages in the steady state are all equal to each other. In other words, the inflow and outflow rates at each stage are equal when the network is in the steady state. Note that the outflow rate of stage k is the inflow rate of stage $k+1$. If the blocking probability of the first stage is π_1, then the flow rate of patients throughout the network is given by:

$$F = \lambda(1 - \pi_1) \tag{4.2}$$

The number of servers and service rate at each stage are adjusted to incorporate the effect of blocking. First, we find the average number of servers at each stage that are blocked by the patients whose services are finished but are not able to move to the next stage. To this end, it is assumed that a virtual waiting line exists before the next stage so that the blocked patients will wait there for a server to become available. The virtual waiting line is assumed to have infinite capacity. The average number of patients waiting in this virtual queue, which is an $M/G/s$ queue, is then determined. For an $M/G/s$ queue with the arrival rate λ and the service rate μ the average length of queue is approximated by (Lee and Longton, 1959):

$$L(M/G/s) \approx \frac{1+C_s^2}{2} L(M/M/s), \tag{4.3}$$

where C_s is the coefficient of variation of service time distribution. The $M/M/s$ queue on the RHS of (4.3) has the same service rate as the $M/G/s$ on the LHS. This approximation is deemed to be an "*excellent*" approximation for $M/G/s$ queues (Whitt, 1993). Note that (4.3) is exact for $M/M/s$ queues since the coefficient of variation of the Exponential distribution is equal to one. The average length of an $M/M/s$ queue is obtained by (Gross et al., 2008):

$$L(\lambda, \mu, s) = \frac{s^s \rho^{s+1}}{s!(1-\rho)^2} \left[\sum_{n=0}^{s-1} \frac{r^n}{n!} + \frac{r^s}{s!(1-\rho)} \right]^{-1}, \tag{4.4}$$

where $r = \lambda/\mu$ and $\rho = r/s$.

In the *Specialization* scenario, all the stages of the hospitals are $M/M/s$ queues. Therefore, Equation (4.4) is used to determine the average number of patients blocked at each stage. However, in the *Diversification* scenario, we deal with $M/H_2/s$ queues

as two types of patients are present in the queues. Hence, the coefficient of variation of the Hyper-exponential distribution is used in (4.3).

Given the arrival rate and the service rate of type-1 patients being λ_1 and μ_1, receptively, and the arrival rate and the service rate of type-2 patients being λ_2 and μ_2, receptively, the mean and the variance of the Hyper-exponential distribution are:

$$\text{Exp(Service Time)} = \frac{\lambda_1(1/\mu_1) + \lambda_2(1/\mu_2)}{\lambda_1 + \lambda_2}, \tag{4.5}$$

and

$$\text{Var(Service Time)} = [\text{Exp(Service Time)}]^2 + 2\frac{\lambda_1\lambda_2}{\lambda_1 + \lambda_2}\left(\frac{1}{\mu_1} - \frac{1}{\mu_2}\right)^2. \tag{4.6}$$

The coefficient of variation is $C_s = \text{Var(Service Time)}^{1/2}/\text{Exp(Service Time)}$.

In reality, this virtual waiting line forms inside the beds of the current stage. So the average number of blocked servers in a stage is the minimum of the average number of waiting patients for the next stage and the total number of servers at this stage. For example, for stage k, the average number of blocked servers is $\min\{s_k, L_{k+1}\}$. The number of effective servers is defined as the total number of servers minus those that are blocked if this value is positive and zero otherwise, which is

$$s_k^* = [s_k - L_{k+1}]^+. \tag{4.7}$$

We now approximate the actual time that patients spend at each stage. The patient's real length of stay at stage k is the original service time, which is on average $\frac{1}{\mu_k}$, plus the time they wait until one server becomes available at stage $k+1$, which is on average $\frac{1}{s_{k+1}\mu_{k+1}}$, weighted by the number of effective servers and the number

of blocked servers, respectively. Therefore, the effective service rate at stage k is defined as:

$$\mu_k^* = \left[\frac{s_k^*}{s_k} \frac{1}{\mu_k} + \frac{s_k - s_k^*}{s_k} \frac{1}{s_{k+1}\mu_{k+1}} \right]^{-1}. \tag{4.8}$$

Using (4.1)-(4.8), the blocking probability at the first stage can be estimated through the following algorithm:

Heuristic Proposed by Bretthauer et al. (2011)

1. Set $m = 0$, $\pi_1^0 = 0$, $\mu_k^0 = \mu_k$, $s_k^0 = s_k$ for $k = 1, \cdots, n$.
2. Increase m by one.
3. Use (4.2) to find the flow rate $F^m = \lambda(1 - \pi_{m-1})$.
4. For stages $k = 1, \cdots, n$, use Equations 4.4-4.8 to update the number of effective servers; $s_k^m = \left[s_k - L(F^m, \mu_{k+1}^{m-1}, s_{k+1}^{m-1}) \right]^+$ and the effective service rates; $\mu_k^m = \left[\frac{s_k^m}{s_k} \frac{1}{\mu_k} + \frac{s_k - s_k^m}{s_k} \frac{1}{s_{k+1}\mu_{k+1}} \right]^{-1}$.
5. Use (4.1) to update the blocking probability; $\pi_1^m = \pi(F^m, \mu_1^m, s_1^m)$.
6. If $\left| \pi_1^m - \pi_1^{m-1} \right| > \delta$, repeat steps 2-5.

4.6. Characterization of a Multi-Hospital Network

Each network of hospitals is described by a set of main parameters (Λ, $\mathbf{M}$, and B_k) as explained in section 4.4. The objective of this section is to substitute this set of parameters with another set of parameters that will be more helpful in interpreting the results of this study. There will be a one-to-one relationship between the two sets of parameters such that for any given values of parameters in one set the values of parameters in the other set can be uniquely determined. The new set of parameters will be used in developing a heuristic algorithm for the **MSBA** problem in Section 4.7 as well as in designing a comprehensive experimental study in Section 4.8.

Without loss of generality, only two types of patients (type-1 and type-2 patients) and two hospitals (Hospital 1 and Hospital 2) are considered to be present in the network. The parameters will be explained in three levels: (1) patient level; (2) hospital level; and (3) network level.

4.6.1 Patient Level Parameters

The parameters that are introduced in this section are associated with the LOS of patients. Let us define θ_k as the ratio of type-1 patient's service rate to type-2 patient's service rate at stage k: $\theta_k = \frac{\mu_{1k}}{\mu_{2k}}$. This parameter indicates the relative average LOS of patients at each stage (Figure 4–3).

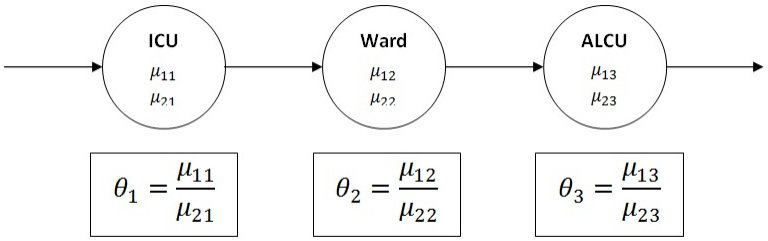

Figure 4–3: Ratio of service rates of patients at each stage

For each patient type, we are also interested in the ratio of service rate at the first stage to stage k's service rate: $\kappa_i^{1k} = \frac{\mu_{i1}}{\mu_{ik}}; k \neq 1$. This parameter describes the relative LOS at each unit of hospital for a given patient type.

For a problem with two patient types and two hospitals, there will be seven parameters and six service rates. If the service rate of type-1 patient in stage 1 is normalized to one, then two parameters need to be chosen as dependent parameters to have a system of 5 unknowns (service rates) and 5 equations (one equation for each independent variable). The θ_1, θ_2, θ_3, κ_2^{12}, and κ_2^{13} are selected as independent

variables. Setting values of independent parameters at some pre-determined levels, the service rate matrix is calculated as:

$$\mathbf{M} = \begin{bmatrix} 1 & \frac{\theta_2}{\theta_1 \kappa_2^{12}} & \frac{\theta_3}{\theta_1 \kappa_2^{13}} \\ \frac{1}{\theta_1} & \frac{1}{\theta_1 \kappa_2^{12}} & \frac{1}{\theta_1 \kappa_2^{13}} \end{bmatrix}. \tag{4.9}$$

The values of dependent parameters are calculated accordingly; $\kappa_1^{12} = \frac{\theta 1}{\theta_2}\kappa_2^{12}$, and $\kappa_1^{13} = \frac{\theta 1}{\theta_3}\kappa_2^{13}$.

The average total LOS (tLOS) of type-i patients in the hospital is $\sum_k \frac{1}{\mu_{ik}}; \mu_i = (\sum_k \frac{1}{\mu_{ik}})^{-1}$. Let us define the ratio of tLOS for the two patient types as

$$\theta = \frac{\text{tLOS}_2}{\text{tLOS}_1} = \frac{\mu_1}{\mu_2}. \tag{4.10}$$

The value of θ depends on the service rates obtained by (4.9) and is only used in deriving some other parameters at the level of hospitals.

4.6.2 Hospital Level Parameters

The arrival rate of patients to the network is given by matrix Λ; $\Lambda_{i,j} = \lambda ij$. The total arrival rate to the system is $\lambda = \sum_i \sum_j \lambda_{ij} = \sum_i \lambda_{i\cdot} = \sum_j \lambda_{\cdot j}$. In the *Diversification* scenario, the fraction of patient who arrive at the Hospital 1 is denoted by β ($\beta = \frac{\lambda_{\cdot 1}}{\lambda}$). Consequently, the fraction of patients who go to the Hospital 2 is $1 - \beta$. In the *Specialization* scenario, all type-1 patients will go to the Hospital 1. Denote the fraction of type-1 patients in the network by α ($\alpha = \frac{\lambda_{1\cdot}}{\lambda}$).

In the *Diversification*, scenario the fraction of arrivals to each hospital (β and $1-\beta$) does not completely describe the arrival intensity. It is also important to know

the average tLOS of all patients in each hospital. Therefore, we define

$$\omega = \frac{\frac{\lambda_{11}}{\lambda_{\cdot 1}}\mu_1 + \frac{\lambda_{21}}{\lambda_{\cdot 1}}\mu_2}{\frac{\lambda_{12}}{\lambda_{\cdot 2}}\mu_1 + \frac{\lambda_{22}}{\lambda_{\cdot 2}}\mu_2} = \frac{\lambda_{\cdot 2}}{\lambda_{\cdot 1}}\frac{\lambda_{11}\theta + \lambda_{21}}{\lambda_{12}\theta + \lambda_{22}}. \tag{4.11}$$

Note that in the *Specialization* scenario, the parameter ω is reduced to θ. For given α, β, ω, and the total arrival rate λ, we can find the arrival rate matrix:

$$\Lambda = \begin{bmatrix} \lambda(\alpha+\beta-1) - \lambda_{22} & (1-\beta)\lambda - \lambda_{22} \\ (1-\alpha)\lambda - \lambda_{22} & \lambda_{22} \end{bmatrix}, \tag{4.12}$$

where $\lambda_{22} = \frac{\lambda(1-\beta)(\beta\theta(1-\omega)+(1-\theta)(1-\alpha))}{(1-\theta)(1-\beta(1-\omega))}$.

4.6.3 Network Level Parameters

The service capacity of the network at each stage is affected not only by the service rate of patient mix but also by the total number of beds available at that stage (b_k). The parameter ρ_{1k} is defined as the ratio of the service capacity of the first stage respect to the service capacity of stage k, $k \neq 1$:

$$\rho_{1k} = \frac{b_1[\alpha\mu_{11} + (1-\alpha)\mu_{21}]}{b_k[\alpha\mu_{1k} + (1-\alpha)\mu_{2k}]} = \frac{b_1[\alpha\theta_1 + (1-\alpha)\kappa_2^{1k}]}{b_k[\alpha\theta_k + (1-\alpha)]} \tag{4.13}$$

Note that ρ_{1k} is independent of the network design and is defined at the network level. For given values of ρ_{1k}, we can find the number of beds at each stage that offers the desired relative service capacities. By choosing an arbitrary value for the number of beds at one stage, say b_1, we have for $k \neq 1$:

$$b_k = \frac{b_1[\alpha\theta_1 + (1-\alpha)\kappa_2^{1k}]}{\rho_{1k}[\alpha\theta_k + (1-\alpha)]}. \tag{4.14}$$

4.7. Heuristic for the MSBA Problem

The **MSBA** problem is optimally solved by examining all the possible bed allocations at all the stages. For a two-hospital problem the number of all allocation scenarios at stage k is limited by $B_k - 1$ (at least one bed should be allocated to each hospital, i.e., $b_{jk} \geq 1$, $\sum_j b_{jk} = B_k$, and b_{jk} is integer), and the total number of feasible solutions is $\prod_k (B_k - 1)$. As the number of beds available at each stage (B_k) and/or the number of stages in the clinical pathway of patients increases, the time needed to explore all feasible combinations surges drastically. Therefore, it is important to develop a heuristic algorithm that finds a near-optimal solution for the **MSBA** problem in a reasonable amount of time.

The heuristic proposed in this section is a greedy local search that starts from a very good initial solution. The initial solution is generated using the parameters of the network defined in Section 4.6. In some cases, the generated initial solution matches the optimal solution or is very close to it. However, one should bear in mind that the **MSBA** problem is an integer program with a non-linear objective function. Even though the number of allocated beds in the heuristic solution is sometimes very close to the optimal, the optimality gap could be significant. In the next subsection, this heuristic algorithm is presented.

4.7.1 A Greedy Local Search Heuristic

A local search heuristic is developed to solve the **MSBA** problem in this section. The parameters defined in Section 4.6 are used to calculate some weights associated with the patient loads each hospital handles at each stage. These weights are then used to allocate the beds to the hospitals. The approach explained here is presented

for a problem with two patient types and two hospitals. Nevertheless, it can be easily generalized to a network of more than two hospitals with more than two types of patient.

In the *Diversification* scenario, the fraction of total patients served by Hospital 1 and Hospital 2 is β and $1-\beta$, respectively. However, this does not represent the relative work load of the hospitals as the LOS of patients has not been considered. To account for the LOS of patients in each hospital, a parameter similar to ω (Equation 4.11) is defined for each stage of the network. The weighted average service rates at stage k for the two hospitals are calculated using the arrival rates as weights. Then, the parameter ω_k is defined as the relative LOS of patients between the two hospitals at that stage:

$$\omega_k = \frac{\frac{\lambda_{11}}{\lambda_{\cdot 1}}\mu_{1k} + \frac{\lambda_{21}}{\lambda_{\cdot 1}}\mu_{2k}}{\frac{\lambda_{12}}{\lambda_{\cdot 2}}\mu_{1k} + \frac{\lambda_{22}}{\lambda_{\cdot 2}}\mu_{2k}} = \frac{1-\beta}{\beta}\frac{\lambda_{11}\theta_k + \lambda_{21}}{\lambda_{12}\theta_k + \lambda_{22}} \tag{4.15}$$

The proportion of patients arriving to each hospital is adjusted by ω_k to find the relative patient loads and the beds are allocated according to the patient loads. The allocation weights at stage k for Hospital 1 and Hospital 2 are β and $(1-\beta)\omega_k$, respectively. Using these weights, $\frac{\beta}{\beta+(1-\beta)\omega_k}B_k$ beds are allocated to the stage k of Hospital 1 and the remaining beds, $\frac{(1-\beta)\omega_k}{\beta+(1-\beta)\omega_k}B_k$, are allocated to the stage k of Hospital 2. The number of allocated beds are rounded to the nearest integer values such that their sum is always equal to B_k.

The same approach is applied to find the weights in the *Specialization* scenario. Note that in this scenario $\omega_k = \theta_k$ as each hospital deals with only one patient type. The fraction of total arrivals that enter Hospital 1 and Hospital 2 is α and $1-\alpha$,

respectively. Hence, the adjusted allocation weights at stage k are α and $(1-\alpha)\theta_k$, respectively. The available beds at each stage are distributed according to these weights and the non-integer numbers are rounded to the nearest integer values such that their sum is always equal to B_k. This gives a starting point for the local search algorithm.

Given that the initial solution is our current solution, the beds are shifted from one hospital to the other to improve the quality of current solution. We move only one bed at a time to explore the immediate neighbor of the current solution in the feasible solution space. If this action improves the objective function, we will continue reallocating beds until no further improvement is possible. This is considered as a neighborhood search algorithm.

There are two directions along which we can shift the beds. One direction is adding a bed to the hospital that has fewer beds and subtracting one bed from the other hospital. The other direction is the opposite way; adding a bed to the hospital that has more beds and subtracting one bed from the other hospital. The direction that leads to a larger improvement in the objective function within the first step will be chosen. Thus, we first check in which way the objective function would decrease the most. This implies that the local search has a greedy logic in choosing the direction of search. If none of the two directions improves the objective function, we keep the same allocations of bed and skip to the next stage.

While we are searching for a better solution, each stage is considered separately from others. When no further improvement is possible at a stage, we move to the next stage. The order we choose the stages is determined as follows. The service

capacity of each stage at the network level is defined using a parameter similar to Equation 4.13. The service capacity of stage k is computed as

$$\rho_k = b_k[\alpha\mu_{1k} + (1-\alpha)\mu_{2k}]. \tag{4.16}$$

The congestion of stage k is then $1/\rho_k$. We start from the stage that has the highest congestion level and move towards the stage with the lowest congestion level in the network. The Greedy Neighborhood Search (**GNS**) algorithm is summarized in the Table 4–4.

4.8. Design of Experiment

To answer the main research question of this study, we are interested in finding the network settings in which narrowing down the scope of care, i.e., *Specialization*, is a superior strategy over broadening the scope of care, i.e., *Diversification*. This requires examining all possible settings of the network and evaluating the performance of both scenarios in each setting. In this section, an experiment including numerous problem instances is designed through considering meaningful values for the network parameters that were defined in section 4.6.

4.8.1 Problem Instances

In choosing the values of network parameters for designing the experiment, all the potential settings for a network of hospitals are envisaged. The values of interest for each parameter that are considered are reported in Table 4–5.

We now elaborate on the rationale of choosing these values in our experiment. No value less than one is chosen for θ_k. This means that the average LOS of type-1 patients in never greater than that of type-2 patients in any stage. By doing so, we

Table 4–4: The heuristic algorithm for the **MSBA** problem

The GNS Heuristic Bed Allocation Algorithm

Initialization Phase:
Step 1: For given matrices of Λ and $\mathbf{M}$, calculate α, β, θ_k, and ω_k.
Step 2: Compute the allocation weights at each stage based on the network parameters:

(i) *Diversification* scenario: the weights at stage k for Hospital 1 and Hospital 2 are respectively β and $(1-\beta)\omega_k$.

(ii) *Specialization* scenario: the weights at stage k for Hospital 1 and Hospital 2 are respectively are α and $(1-\alpha)\theta_k$.

Step 3: Allocate the available beds at stage k (B_k) to each hospital:

(i) *Diversification* scenario: the allocated beds at stage k to Hospital 1 and Hospital 2 are respectively $\frac{\beta}{\beta+(1-\beta)\omega_k}B_k$ and $\frac{(1-\beta)\omega_k}{\beta+(1-\beta)\omega_k}B_k$ (round the non-integer number of beds).

(ii) *Specialization* scenario: the allocated beds at stage k to Hospital 1 and Hospital 2 are respectively $\frac{\alpha}{\alpha+(1-\alpha)\theta_k}B_k$ and $\frac{(1-\alpha)\theta_k}{\alpha+(1-\alpha)\theta_k}B_k$ (round the non-integer number of beds).

Step 4: Find the maximum blocking probability of two hospitals using the solution of Step 3.

Improvement Phase:
Step 5: Compute the congestion level at each stage at the network level (Congestion Level: $1/\rho_k$).
Step 6: Iterate over all stages (from the highest congested to the lowest congested stage)

Step 6.1: Find the direction along which moving one bed from one hospital to the other decreases the objective function the most. If no direction is found, skip to the next stage.

Step 6.2: Reallocate the beds in the direction found in the step 6.1 one at a time until no improvement is possible.

Table 4–5: Parameters values in the experimental study

Parameter	Values	of	Interest
α	0.25	0.50	0.75
β	0.25	0.50	
ω	0.8	1.0	1.2
$\theta_k, k = 1, 2, 3$	1	2	5
$\kappa_2^{1k}, k = 2, 3$	0.5	1	2
$\rho_{1k}, k = 2, 3$	0.5	1	2
λ	1	2	

recognize the type-1 patients as *fast* patients and type-2 patients as *slow* patients. Therefore, the parameter α will represent the fraction of fast patients in the system. The parameter β does not have any value greater than half since the two hospitals are identical. The parameters κ_2^{1k} and ρ_{1k} are assigned values less than, equal to, and greater than one to capture all the possible situations. The parameter ω also has different values to simulate different average tLOS in each hospital.

The number of all combinations produced from setting the parameters at the values of Table 4–5 is 78,732 problem instances. Note that some combinations are not feasible. Thus, this will be the maximum number of problem instances that are considered in the experiment. The heuristic algorithm developed in section 4.7 makes this comprehensive study with this large number of problem instances doable. This heuristic algorithm finds a good solution for the **MSBA** problem in a reasonable amount of time while the optimal solution is very time-consuming to obtain through complete enumeration.

4.8.2 Outputs of the Experiment

The aim of this study is to delineate the settings in which the *Specialization* scenario is a better configuration for a healthcare network. The performance measurement considered in this study is the maximum blocking probability of two hospitals, which has been defined in the objective function of the **MSBA** problem. However, a large number of problem instances is considered in the experimental study and we are interested in identifying those cases where narrowing down the scope of services is beneficial. Therefore, the following outputs are reported in the experimental study:

Output 1: Possibility of Improvement. This output is calculated based on the percentage of problem instances in which the *Specialization* decreases the maximum blocking probability of the network. It reflects the possibility of improvement at the level of network through *Specialization* of services.

Output 2: Impact. Average reduction in the maximum blocking probability in those problem instances where *Specialization* is preferred will be reported by Output 2. This shows the real impact of *Specialization* on the performance of the network.

Output 3: Acceptability. This output shows the percentage of problem instances in which both hospitals benefit from the *Specialization* among those cases identified in Output 1 . It describes the acceptability of the *Specialization* by all the hospitals involved in the restructuring process and captures the possibility of improvement at the level of individual hospital.

Cross-sectional analysis of the results is conducted to examine the three outputs defined above from different angles and facilitate the interpretation of results. In Output 1, we look into the fraction of original settings of the problem where the

whole network is better off by the *Specialization* of services. Note that we compare the solution of best bed allocation in each scenario to that of another scenario. This might enforce shifting some resources from one hospital to another to achieve the best performance of the network. Therefore, we also search for cases in which both hospitals are better off by the *Specialization*, which is captured in Output 3. We also report on the average performance improvement of the network by the *Specialization* in Output 2.

4.8.3 Performance Evaluation of the GNS Heuristic

Before turning to presenting the results, the performance of the proposed heuristic algorithm for the **MSBA** problem is evaluated through comparing its solution with the optimal solution, which is found through complete enumeration. Only small-size problem instances are considered for this comparison where the optimal solution can be found in a reasonable amount of time. The problem instances are generated using the same parameters of Table 4–5. But the comparison is restricted to those problem instances with the total number of beds in all stages being less than 30. The gap between the objective function values associated with the solution of the heuristic algorithm and complete enumeration as well as the average computation time of each approach are reported in Table 4–6.

Table 4–6: Performance evaluation of the heuristic for the **MSBA** problem

	Optimality Gap	Computation	Time (sec)
	(%)	Optimal	Heuristic
All Problems	3.00	674.74	3.41
Diversification	4.20	709.07	3.80
Specialization	0.35	598.22	2.52

The error associated with the heuristic algorithm is very small (3%) in all the cases that have been considered. In the *Specialization* cases, the heuristic algorithm seems to perform even better and generate solutions very close to the optimal ones. In terms of computation time, the heuristic algorithm saves a considerable amount of time when being used instead of the complete enumeration approach.

4.9. Specialization vs. Diversification

In this section, the results obtained from the experimental study of Section 4.8 are analyzed to provide insights to the problem of *Diversification* versus *Specialization* of healthcare services. As explained earlier, a wide range of parameters of a multi-hospital network has been considered to generate a large number of problem instances (approximately 60,000 instances). Through comparing the performance of two different scenarios in terms of blocking probability of patients the situations in which *Specialization* is a superior scenario over *Diversification* is characterized.

The parameters introduced in Section 4.6 include all dimensions of a healthcare network. However, the analysis is focused on a few important characteristics of the network to establish some rules to lay out the situations where each scenario dominates the other one. The first characteristic is the fraction of arrivals associated with the patients whose LOS in the hospital is less than the other type (so-called fast patients). In the designed experiment it was assumed that the LOS of type-1 patients at each stage is always less than that of type-2 patients. The fraction of type-1 patients to the all patients entering to the network is denoted by α.

Although it is important to know about the mix in the arrival of slow and fast patients to a network, the magnitude of difference between their tLOS is also a

significant factor. This relative difference is indicated by θ (Equation 4.10). As θ increases, the role of α becomes more substantial as well. The other factor of interest is the relative tLOS of patients being served in the two hospital, which is expressed by ω (Equation 4.11). Note that θ and ω reflect the ratio of tLOS between the two hospitals in the *Specialization* and *Diversification* scenarios, respectively.

In terms of patient load that two hospitals receive in each scenario, we must incorporate the arrival rate of each type. To this end, the proportion of patients arriving at each hospital is adjusted by their tLOS similar to the approach used in developing the heuristic algorithm for the **MSBA** problem in section 4.7. In the *Diversification* scenario, the patient load for Hospital 1 is β and for Hospital 2 is $(1-\beta)\omega$. The relative patient loads are defined as $d = \frac{\beta}{(1-\beta)\omega}$. In the *Specialization* scenario, the relative patient load is defined as $s = \frac{\alpha}{(1-\alpha)\theta}$.

The results are presented through Figure 4–4 to Figure 4–9 in this section. Each figure shows the outputs of the experiments from a perspective of one parameter of interest. The observations associated with these figures are summarized in the form of some insights to the problem of multi-hospital network configuration.

4.9.1 Possibility of Improvement by Specialization

In the experiments, the Output 1 shows the percentage of problem instances in which the *Specialization* is recommended. The following observations are made:

1. Overall, in 32.27% of all cases, narrowing down the scope of care is the superior option.

2. As the percentage of fast patients (α) increases in the network, it is less likely that the *Specialization* improves the performance of the system. (Figure 4–4-Output

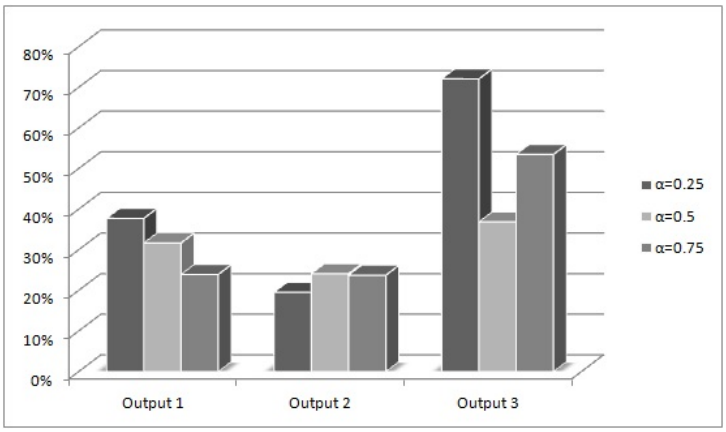

Figure 4–4: Results of experimental study - all outputs in terms of α

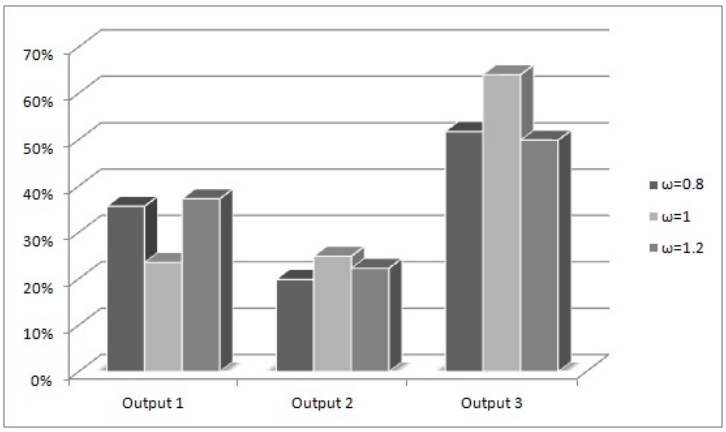

Figure 4–5: Results of experimental study - all outputs in terms of ω

1). This implies that when the majority of patients have short LOS in the hospital we are better off by dispersing them over the network rather than allocating all of them to only one hospital.

3. We see that the possibility of improvement by the *Specialization* is larger when the tLOS of patients in two hospitals before *Specialization* is not equal, i.e., ω =0.8 or 1.2 (Figure 4–5-Output 1). When omega=1, then output 1 values is significantly smaller.

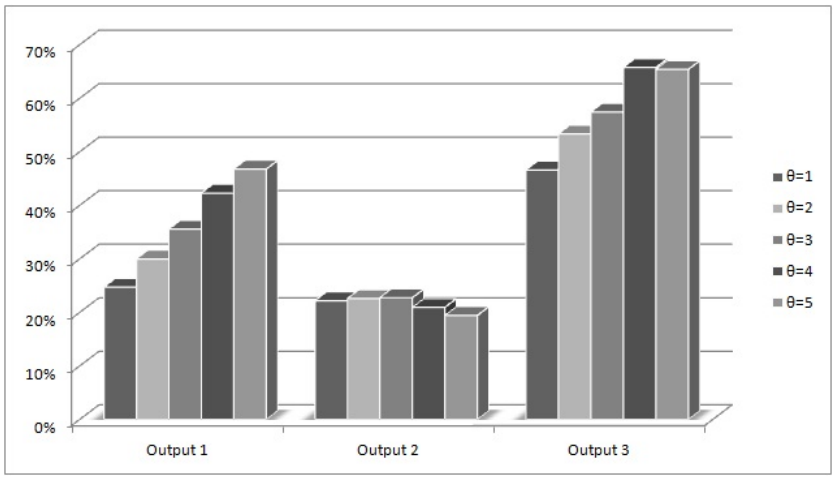

Figure 4–6: Results of experimental study - all outputs in terms of θ

4. In general, as the tLOS ratio of slow to fast patients in two hospitals after *Specialization* (θ) increases, the chance of improvement seems to increase as well (Figure 4–6-Output 1). This means that if the flow of patients of one type is much slower than the other, it would be better to keep them separated from each other.

5. We observe a greater chance of improving through the *Specialization* in cases with unequal patient loads ($d \neq 1$) (Figure 4–7). It seems that when the patient load is balanced between the two hospitals, we will gain no benefits by the *Specialization* in most cases.

4.9.2 Impact of Specialization

We express the average performance improvement the *Specialization* can make over the *Diversification* scenario with the Output 2. The observations regarding this output are:

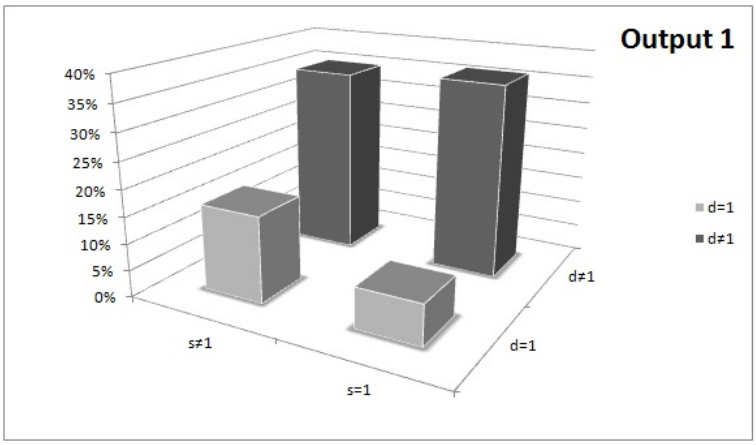

Figure 4–7: Results of experimental study - Output 1 in terms of patient load

1. Overall, the average performance improvement in those cases where the *Specialization* dominates the *Diversification* is 22.21% decrease in the maximum blocking probability of two sites.

2. The average improvement seems to be robust respect to the value of α (Figure 4–4-Output 2). Interestingly, this holds true for other parameters; ω, θ, d, and s (Figure 4–5-Output 2, Figure 4–6-Output 2, and Figure 4–8). In all cases, regardless of the network parameters, the percent decrease in blocking probability obtained through the *Specialization* of services is around 20%.

3. However, there is a slight increase in the Output 2 when the majority of patients in the network are fast patients ($\alpha = 0.75$), or when the service rate of patients in two hospitals are around the same range before and after the *Specialization* (θ is small and $\omega = 1$), or when the patient loads are unbalanced in the *Diversification* scenario.

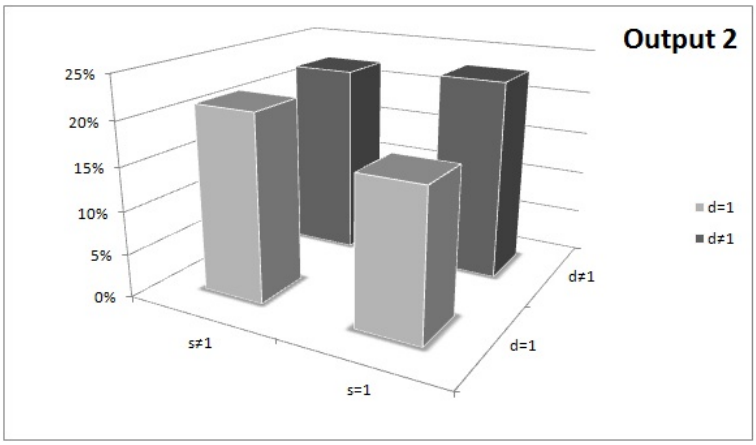

Figure 4–8: Results of experimental study - Output 2 in terms of patient load

4.9.3 Acceptability of Specialization

The acceptability of change in the network configuration by the two sites is indicated by the Output 3. We can claim that:

1. Overall, in 54.78% of cases in which *Specialization* outperforms *Diversification* the blocking probability in two hospitals decreases simultaneously.

2. It seems that when there is some skewness in the arrival patient mix (α=0.25 or α=0.75), it is more likely that both hospital will benefit from *Specialization*. The chance of bilateral improvement is the highest when α=0.25. This contradicts the expectation that when one hospital is supposed to take care of small number of patients who are all fast patients, the other hospital will refuse such change. But, counter-intuitively, performance of both hospitals will be enhanced by this effort (Figure 4–4-Output 3).

3. Also, as the gap between the tLOS of patients becomes larger, both hospitals will be satisfied in more cases (Figure 4–6-Output 3).

4. When the patient load is evenly distributed between the two sites ($d = 1$) there will be a higher chance that the two hospitals embrace *Specialization* (Figure 4–9). This is true even when the outcome is unevenly distribution of patient loads ($s \neq 1$). When the hospitals are currently handling unequal loads of patients ($d \neq 1$) it will be more challenging to convince both sides to switch from *Diversification* to *Specialization*.

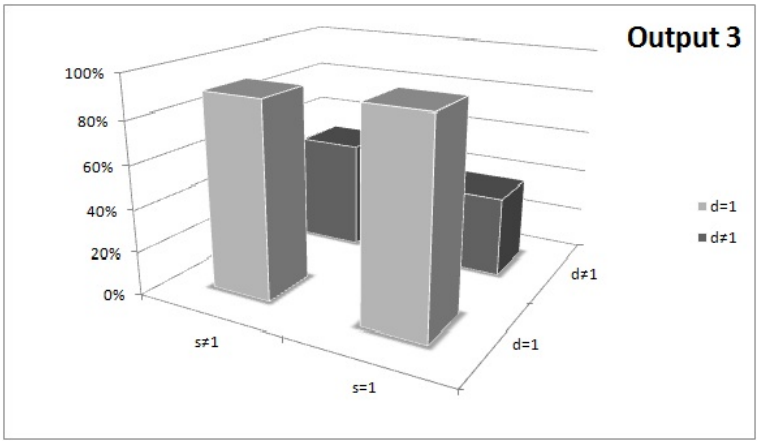

Figure 4–9: Results of experimental Study - Output 3 in terms of patient load

4.10. Conclusion

This chapter examines the admission process of patients in a multi-hospital network and considers two system configuration scenarios. Specifically, it compares the performances of *Specialization* and *Diversification* options and lays out the situations in which each scenario outperforms the other. Moreover, the bed allocation problem for a network of hospitals has been formulated and an efficient heuristic algorithm has been presented. A comprehensive experimental study was designed and used

to evaluate the scenarios in terms of refusal rate of patients at the hospitals of the network.

The observations derived from the experimental study are summarized in Figure 4–10 and Table 4–7. The main insights are: (i) the possibility of improvement through *Specialization* decreases as the percentage of patients with short LOS entering the network increases or the difference between LOS of different types of patients decreases or the patient load between the hospitals is balanced; (ii) the impact of *Specialization* is quite robust respect to the network parameters; and (iii) the mutual improvement of *Specialization* happens when there is asymmetry in the patients' arrivals or LOSs handled by each hospital or symmetry in the patient loads of hospitals before *Specialization*.

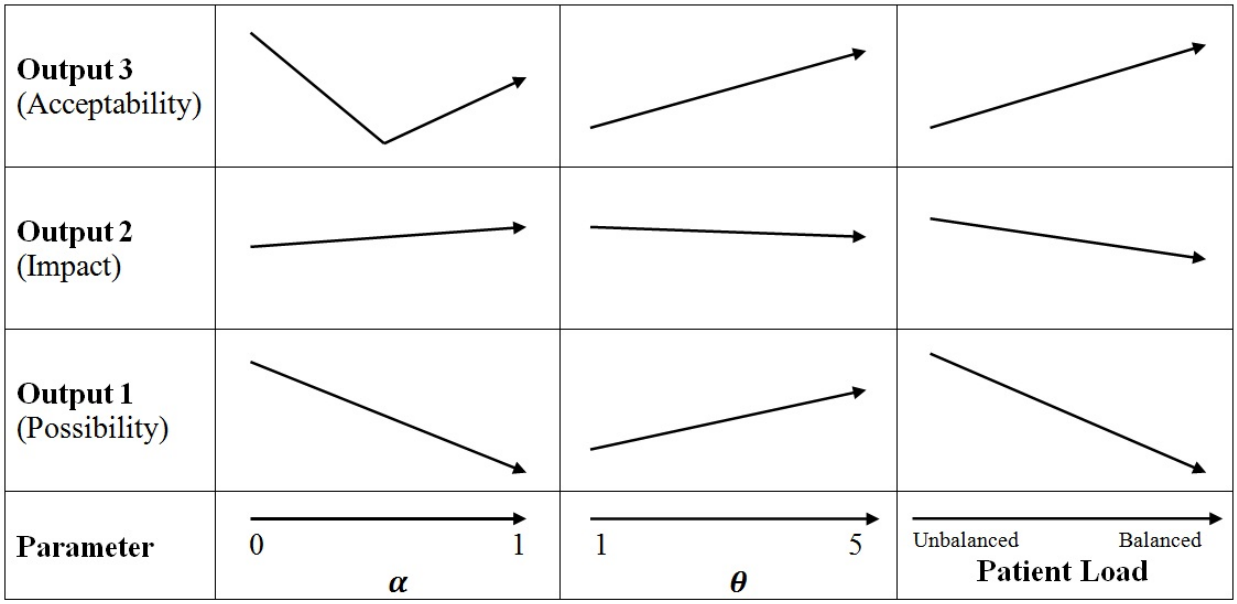

Figure 4–10: Summary of insights from the experimental study

Table 4–7: Effects of network parameters on the exercise of specialization

Parameter	Percentage of Fast Patients	Difference in LOS	Patient Load
Possibility	decreases as α increases	increases as θ increases	decreases as the load is more balanced
Impact	slightly increases as α increases	slightly decreases as θ increases	decreases as the load is more balanced
Acceptability	disproportionately decreases as α increases	increases as θ increases	increases as the load is more balanced

The results obtained in the simulation study of restructuring the MNH and the MGH stroke wards in Section 4.2.2 is consistent with the insights provided by the experimental study. The data used for the simulation study gives the following parameters: $\alpha = 0.7$, $\theta = 1.91$, $\omega = 1$, $g = 1$, and $s = 1.22$. These values suggest that the *Specialization* is not an appropriate recommendation for organizing the stroke wards at the MNH and the MGH, which is confirmed by the results of simulation study.

As more structural changes in the healthcare networks are triggered by the significant budget cuts or by the desire to better streamline the healthcare processes, it is important to review benefits of any proposed change before putting it into action. Therefore, the health authorities who have significant impact on such decisions (e.g., L'Agence in Montreal and Ministère de la Santé et des Services Sociaux in Quebec) are recommended to take advantage of analytical frameworks, such as the one presented in this chapter, rather than the top-down and intuition-based approaches they often adopt for policy design.

In the modeling framework of this chapter, tandem queues were used to describe the flow of stroke patients in hospital. While this holds true for the majority of

patients, a small fraction of patients might revisit the ICU or other units more than once during their stay in the hospital. A more general model that incorporates flexible patients routing would be a more realistic and valuable modification of this study.

In this study, a simple network of two hospitals with two types of patients was considered. A useful extension of this analysis includes validation of the results for a problem with more than two hospitals and two types of patient. While most components of this study can be easily revised to accommodate such an extension, the validation of the results seems unavoidable. Further, the problem of designing the network and finding the optimal level of flexibility at each site (that determines which services should be offered in each hospital) is an interesting direction for future research.

CHAPTER 5
Concluding Remarks and Future Research

This thesis studies capacity-related policies regarding inpatient beds in acute care wards recognizing multiple types of patients with different medical characteristics. The financial pressures in the healthcare sector, coupled with the drive for improving quality and efficiency, have exposed hospital managers to various challenges in developing alteration plans for their hospitals. A key factor in improving operations of healthcare systems is better use of available resources, such as inpatient beds. The Operation Research models, which are developed around the resource management issues, could be used by decision makers in healthcare to assist them with making the best capacity decision concerning resources. Two applications of such models to managerial problems in a neurological hospital were illustrated in this thesis. In particular, the patient admission policy design for a neurology ward (tactical decisions) and network configuration and bed allocation for stroke wards of two hospitals (strategic decisions) have been addressed.

5.1. Summary of Research Findings

In Chapter 2 of this thesis, a literature review of the patient admission and bed allocation problem was presented that is based on a thorough survey of the state of the art in this domain. The previous studies have been categorized as bed allocation problems at the level of hospitals located in one region, at the level of wards (units) of a hospital, and at the level of patient types within a ward of a hospital. The research

related to patient admission problem was introduced in a separate category. This Chapter is believed to have the potential to be a segment of a more comprehensive review paper in this domain.

In Chapter 3, an admission control and bed allocation problem in a neurology ward was studied. Many neurology wards face the problem of insufficient capacity to meet demand for inpatient beds, especially during demand surges. The problem is pronounced since admitting these patients to other wards is not an option, i.e., off-unit servicing is not feasible for neurology patients. Recent studies have shown that neurology patients are more effectively treated in specialized neurology wards that offer properly organized care. To the best of the author's knowledge, this is the first study that makes an explicit effort to model the differentiating features of neurology wards, and hence provides managerial insights specific to this domain.

The process of admitting neurology patients from ED to the ward and transferring them to another hospital was modeled using an infinite-horizon average cost dynamic program. None of the beds in the ward were dedicated to certain patient types in this model. It was shown that the optimal policy for admitting patients from the ED is dynamic and depends on the state of the system. To solve the average cost DP and overcome the curse of dimensionality, an LP-based approximate dynamic programming (ADP) approach was developed that uses some information from a static model. While this method typically involves a large-scale LP, our approach involves solving a number of small DPs that are derived by employing a non-linear functional approximation. To the best of the author's knowledge, the ADP for the average cost problem has not been fully explored theoretically.

In many neurology wards (including the hospital considered in this study) a static patient admission policy is used by assigning a fixed number of beds to each type of disease. Also, transferring patients is triggered after a certain amount of ED boarding time. In contrast, the proposed ADP policy does not use earmarked beds and decides to transfer the patient at the time of arrival, considering the state of the system. The comparative analysis showed that the ADP policy decreases the average boarding time in the ED (especially when there are is a limited number of beds available or patient's sensitivity to waiting is high) significantly without affecting the average rate of patient transfers. This translates to smaller deterioration in health status of patients resulting from waiting in ED and delays in receiving the specialized care provided in the neurology ward. Acknowledging the challenges in implementing the fully state dependent ADP policy, a Priority Cut-off policy based on the ADP was developed that performs quite well. This policy follows the ADP policy in the critical states of the system and in other states applies FCFS policy.

The main insight for neurology managers is that it would be better to decide on the admission or transferring a patient based on the state of the system. Also, an active decision making approach on transfers that acts upon arrival of patients outperforms other passive procedures that wait for a period of time and then request for transferring the patient. It is recommended to keep the beds flexible to serve all types of patients instead of dedicating them to patient types. If the managers prefer to use an earmarking strategy, it is suggested to do so based on the level of severity of the patients' condition rather than their disease (i.e., along the lines of the PC policy). The level of severity largely determines the patients' LOS of patients or

the impact of unavailability of resources on patient's health status. Therefore, the allocation policies according to the categorization that is based on level of severity are expected to perform better.

In Chapter 4 of this thesis, the problem of network configuration and multi-site resource allocation was considered. This research has been inspired by a structural change in the administration of two stroke wards in Montreal. The possible scenarios for managing a network of multiple acute care wards that provide multiple levels of care are specialization and diversification. In the specialization scenario, each ward is dedicated to providing care to specific type(s) of patients. In contrast, the diversification scenario requires every ward to accommodate all types of patients.

While a significant number of papers (in healthcare and other areas) are found in the literature that study such a problem, there are some important features specific to this research. In most previous works, specialization of service refers to multiple queues that have separate waiting lines and the alternative option is a pooled system with one single queue that feeds all servers. Even though in this thesis the specialization scenario is similar, the diversification of services represents centers that are managed independently. Each center has its own patient mix, resources, and demand. Furthermore, the performances of hospitals in each scenario are evaluated using a tandem queueing network that incorporates blocking effects in patients' clinical path. From a modeling perspective, a multi-site bed allocation problem was presented so that the resources are optimally distributed in accordance with the patients flow in each hospital.

To characterize the situations in which one scenario dominates the other, a wide range of network parameters has been considered. These parameters were also used to construct a heuristic algorithm to solve the multi-site bed allocation problem. By considering a comprehensive set of problem instances, the performances of both scenarios have been thoroughly examined. The results show that only in one third of cases specialization improves the performance of systems and the blocking probability of network is reduced by one fifth in those cases.

Another important factor of this research is the consideration of mutual improvement for all the hospitals of a network. It is very common in healthcare systems, especially in publicly funded ones where financial incentives are absent, that hospital managers resist structural changes if no benefit is associated with them. The study looked into this issue and observed that in around half of the cases that specialization advances the performance of the overall network all the sites will be better off by this change.

The other observations of this study are helpful in evaluating outcomes of narrowing down the scope of healthcare services in a multi-hospital network that currently has a general configuration. Based on the properties of the network, the possibility, impact, and acceptability of specialization can be determined. As presence of fast patients respect to patients with longer LOS in the network increases it is more possible that the specialization is beneficial. Meanwhile, when the difference between the LOSs of patients is multiplied the possibility of improvement also increases. But if the patient load is unequally distributed between the hospitals the chance of improvement declines.

The impact of improvement by specialization is somewhat robust to the network properties. However, slight increase and decrease are observed when the majority of patients have shorter LOS and difference in the LOSs expands, respectively. The acceptability of change by hospitals also depends on the characteristics of the network. The interesting insight is that the acceptability is more often achieved when there is asymmetry in the patient mix arrivals or their LOSs. But if we look at the patient load, which reflects both the arrival and LOS distributions, the symmetry of patient loads between the hospitals in the diversified scenario implies more cooperation from all sites, which is very intuitive.

5.2. Future Research Directions

A future research direction of this thesis is studying admission policies and designing transfer protocols after a merger of two hospitals takes place. This idea is blossomed from the two pieces that have been studied in this thesis. Recently, more changes in managing the acute care wards are happening and hospitals are moving toward partnerships with other hospitals to collectively provide care to patients. Therefore, designing patient admission and transfer policies for more than one acute care ward appears to be a valuable research endeavor.

The complexity of decision making process in such situations originates from the fact that the state of other acute care wards must be taken into account. This issue is missing in the patient admission policy design considered in this thesis, in which it was assumed that transfer of patients is possible at all times. This essentially implies that we consider a fixed pre-known waiting time (and consequently a fixed transfer cost) in the destination ward. However, this cost is dynamic and changes over time

according to the congestion level of the other ward. A modeling framework that incorporates this important feature will provide more realistic insights.

The stationary assumption regarding the arrival process is a substantial limitation of this study. Some modeling components included in this thesis can be revised to adopt time-varying arrival rates. However, another future research is to incorporate the non-stationary arrivals into the models and observe the sensitivity of results to this change. The problem of dynamic allocation of inpatient beds needs to be formulated to address the variations of resource requirements over time.

5.3. Concluding Remarks

The setting of neurology wards constituted the basic platform for modeling the problems studied in this thesis. The calibration of models and cases studies were also conducted using the data collected from neurology inpatients. However, the theoretical frameworks and the results can be generalized to other acute care wards where multiple types of patients are competing for common resources and where providing timely access to care is the first priority. Moreover, it is emphasized that the implications of this research is not confined to the healthcare domain. The contributions and insights can be cautiously extended to other admission control problems in multi-class queueing systems.

In closing, this thesis provides insights for the problem of capacity allocation policies for inpatient beds in acute care wards. Hospital managers and healthcare authorities can use the results of this research to plan and implement fundamental changes in their systems and enhance the operational efficiency and quality in their

healthcare institutions. This research is expected to build trust in healthcare policy makers to deploy more analytical frameworks and optimization models in their improvement efforts.

REFERENCES

Abadi, IN Kamal, Nicholas G Hall, Chelliah Sriskandarajah. 2000. Minimizing cycle time in a blocking flowshop. *Operations Research* **48**(1) 177–180.

Abdelaziz, F Ben, M Masmoudi. 2011. A multiobjective stochastic program for hospital bed planning. *Journal of the Operational Research Society* **63**(4) 530–538.

Akcali, Elif, Murray J Coté, Chin Lin. 2006. A network flow approach to optimizing hospital bed capacity decisions. *Health Care Management Science* **9**(4) 391–404.

Altiok, Tayfur, Harry G Perros. 1987. Approximate analysis of arbitrary configurations of open queueing networks with blocking. *Annals of Operations Research* **9**(1) 481–509.

Ata, Baris, Jan A Van Mieghem. 2009. The value of partial resource pooling: Should a service network be integrated or product-focused? *Management Science* **55**(1) 115–131.

Ayvaz, N., W.T. Huh. 2010. Allocation of hospital capacity to multiple types of patients. *Journal of Revenue & Pricing Management* **9**(5) 386–398.

Balsamo, Simonetta, Vittoria de Nitto Personé, Raif Onvural. 2001. *Analysis of queueing networks with blocking*, vol. 31. Springer.

Bekker, R, AM de Bruin. 2010. Time-dependent analysis for refused admissions in clinical wards. *Annals of Operations Research* **178**(1) 45–65.

Benjaafar, Saifallah. 1995. Performance bounds for the effectiveness of pooling in multiprocessing systems. *European Journal of Operational Research* **87**(2) 375–388.

Bertsekas, Dimitri P. 2005. *Dynamic programming and optimal control*. Athena Scientific.

Brandwajn, Alexandre, Yung-Li Lily Jow. 1988. An approximation method for tandem queues with blocking. *Operations Research* **36**(1) 73–83.

Bretthauer, Kurt M, H Sebastian Heese, Hubert Pun, Edwin Coe. 2011. Blocking in healthcare operations: A new heuristic and an application. *Production and Operations Management* **20**(3) 375–391.

Buzacott, John A. 1996. Commonalities in reengineered business processes: models and issues. *Management Science* **42**(5) 768–782.

Castillo, Jose. 1999. Deteriorating stroke: Diagnostic criteria, predictors, mechanisms and treatment. *Cerebrovasc Disease* **9**(suppl 3) 1–8.

Chalfin, Donald B., Stephen Trzeciak, Antonios Likourezos Brigitte M. Baumann, R Phillip Dellinger. 2007. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Critical Care Medicine* **35**(6) 1477–1483.

Cochran, JK, K Roche. 2007. A queuing-based decision support methodology to estimate hospital inpatient bed demand. *Journal of the Operational Research Society* **59**(11) 1471–1482.

Dallery, Yves, Yannick Frein. 1993. On decomposition methods for tandem queueing networks with blocking. *Operations Research* **41**(2) 386–399.

de Farias, Daniela Pucci, Benjamin Van Roy. 2006. A cost-shaping linear program for average-cost approximate dynamic programming with performance guarantees. *Mathematics of Operations Research* **31**(3) 597–620.

Demeester, Peter, Wouter Souffriau, Patrick De Causmaecker, Greet Vanden Berghe. 2010. A hybrid tabu search algorithm for automatically assigning patients to beds. *Artificial intelligence in medicine* **48**(1) 61–70.

Dijk, Nico M, Erik Sluis. 2008. To pool or not to pool in call centers. *Production and Operations Management* **17**(3) 296–305.

El-Darzi, E, C Vasilakis, T Chaussalet, PH Millard. 1998. A simulation modelling approach to evaluating length of stay, occupancy, emptiness and bed blocking in a hospital geriatric department. *Health Care Management Science* **1**(2) 143–149.

Esogbue, A.O., A.J. Singh. 1976. A stochastic model for an optimal priority bed distribution problem in a hospital ward. *Operations Research* **24**(5) 884–898.

Garg, Lalit, Sally McClean, Brian Meenan, Peter Millard. 2010. A non-homogeneous discrete time markov model for admission scheduling and resource planning in a cost or capacity constrained healthcare system. *Health care management science* **13**(2) 155–169.

Gershwin, Stanley B. 1987. An efficient decomposition method for the approximate evaluation of tandem queues with finite storage space and blocking. *Operations Research* **35**(2) 291–305.

Gorunescu, Florin, Sally I McClean, Peter H Millard. 2002a. A queueing model for bed-occupancy management and planning of hospitals. *Journal of the Operational Research Society* 19–24.

Gorunescu, Florin, Sally I McClean, Peter H Millard. 2002b. Using a queueing model to help plan bed allocation in a department of geriatric medicine. *Health Care Management Science* **5**(4) 307–312.

Green, Linda. 2006. Queueing analysis in healthcare. *Patient flow: reducing delay in healthcare delivery*. Springer, 281–307.

Green, Linda, Peter Kolesar. 1991. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science* **37**(1) 84–97.

Green, Linda V. 2002. How many hospital beds? *Journal Information* **39**(4).

Green, Linda V, Vien Nguyen. 2001. Strategies for cutting hospital beds: the impact on patient service. *Health Services Research* **36**(2) 421.

Green, L.V., S. Savin, B. Wang. 2006. Managing patient service in a diagnostic medical facility. *Operations Research* **54**(1) 11–25.

Gross, Donald, John F. Shortle, James M. Thompson, Carl M. Hartis. 2008. *Fundamentals of queueing theory*. John Wiley and Sons.

Güneş, ED, Hande Yaman. 2010. Health network mergers and hospital re-planning. *The Journal of the Operational Research Society* **61**(2) 275–283.

Harper, PR, AK Shahani. 2002. Modelling for the planning and management of bed capacities in hospitals. *Journal of the Operational Research Society* 11–18.

Helm, J.E., S. AhmadBeygi, M.P. Van Oyen. 2011. Design and analysis of hospital admission control for operational effectiveness. *Production and Operations Management* **20**(3) 359–374.

Hershey, John C, Elliott N Weiss, Morris A Cohen. 1981. A stochastic service network model with application to hospital facilities. *Operations Research* **29**(1) 1–22.

Hillier, Frederick S, Ronald W Boling. 1967. Finite queues in series with exponential or erlang service timesa numerical approach. *Operations Research* **15**(2) 286–303.

Hopman, Wilma M, Jane Verner. 2003. Quality of life during and after inpatient stroke rehabilitation. *Stroke* **34**(3) 801–805.

Hulshof, Peter JH, Nikky Kortbeek, Richard J Boucherie, Erwin W Hans, Piet JM Bakker. 2012. Taxonomic classification of planning decisions in health care: a structured review of the state of the art in or/ms. *Health Systems* **1**(2) 129–175.

Hunt, Gordon C. 1956. Sequential arrays of waiting lines. *Operations Research* **4**(6) 674–683.

Jaracz, Krystyna, W Kozubski. 2003. Quality of life in stroke patients. *Acta Neurologica Scandinavica* **107**(5) 324–329.

Jönsson, Ann-Cathrin, Ingrid Lindgren, Björn Hallström, Bo Norrving, Arne Lindgren. 2005. Determinants of quality of life in stroke survivors and their informal caregivers. *Stroke* **36**(4) 803–808.

Joustra, Paul, Erik Van der Sluis, Nico M Van Dijk. 2010. To pool or not to pool in hospitals: a theoretical and practical comparison for a radiotherapy outpatient department. *Annals of Operations Research* **178**(1) 77–89.

Kao, Edward PC, Grace G Tung. 1981. Bed allocation in a public health care delivery system. *Management Science* **27**(5) 507–520.

Koizumi, Naoru, Eri Kuno, Tony E Smith. 2005. Modeling patient flows using a queuing network with blocking. *Health Care Management Science* **8**(1) 49–60.

Kolesar, P. 1970. A markovian model for hospital admission scheduling. *Management Science* **16**(6) B–384.

Kucukyazici, Beste. 2010. Design and improvement of the care processes for stroke: An analytical approach. Ph.D. thesis, McGill University.

Kucukyazici, Beste, Linda Green, Vedat Verter. 2010. Improving stroke outcomes through operational policies. Extended abstract, Annual Conference of MSOM, Haifa, Israel.

Kusters, Rob J, Petra Groot. 1996. Modelling resource availability in general hospitals design and implementation of a decision support model. *European journal of operational research* **88**(3) 428–445.

Lapierre, S.D., D. Goldsman, R. Cochran, J. DuBow. 1999. Bed allocation techniques based on census data. *Socio-Economic Planning Sciences* **33**(1) 25–38.

Lee, AM, PA Longton. 1959. Queueing processes associated with airline passenger check-in. *Journal of the Operational Research Society* **10**(1) 56–71.

Lee, HS, A Bouhchouch, Y Dallery, Y Frein. 1998. Performance evaluation of open queueing networks with arbitrary configuration and finite buffers. *Annals of Operations Research* **79** 181–206.

Lee, Hyo-Seong, Stephen M Pollock. 1990. Approximation analysis of open acyclic exponential queueing networks with blocking. *Operations Research* **38**(6) 1123–1134.

Li, X., P. Beullens, D. Jones, M. Tamiz. 2009. An integrated queuing and multi-objective bed allocation model with application to a hospital in china. *Journal of the Operational Research Society* **60**(3) 330–338.

Lopez, Alan D, Colin D Mathers, Majid Ezzati, Dean T Jamison, Christopher JL Murray. 2006. *Global burden of disease and risk factors*. Oxford University Press, USA.

Mahar, Stephen, Kurt M Bretthauer, Peter A Salzarulo. 2011. Locating specialized service capacity in a multi-hospital network. *European Journal of Operational Research* **212**(3) 596–605.

Mandelbaum, Avishai, Petar Momčilović, Yulia Tseytlin. 2012a. On fair routing from emergency departments to hospital wards: Qed queues with heterogeneous servers. *Management Science* **58**(7) 1273–1291.

Mandelbaum, Avishai, Petar Momčilović, Yulia Tseytlin. 2012b. On fair routing from emergency departments to hospital wards: Qed queues with heterogeneous servers. *Management Science* **58**(7) 1273–1291.

Mandelbaum, Avishai, Martin I Reiman. 1998. On pooling in queueing networks. *Management Science* **44**(7) 971–981.

Nichols-Larsen, Deborah S, PC Clark, Angelique Zeringue, Arlene Greenspan, Sarah Blanton. 2005. Factors influencing stroke survivors quality of life during subacute recovery. *Stroke* **36**(7) 1480–1484.

Osorio, Carolina, Michel Bierlaire. 2009. An analytic finite capacity queueing network model capturing the propagation of congestion and blocking. *European Journal of Operational Research* **196**(3) 996–1007.

Patrick, J., M.L. Puterman, M. Queyranne. 2008. Dynamic multipriority patient scheduling for a diagnostic resource. *Operations Research* **56**(6) 1507–1525.

Perros, Harry G. 1994. *Queueing networks with blocking*. Oxford University Press, Inc.

Perros, Harry G., Tayfur Altiok. 1986. Approximate analysis of open networks of queues with blocking: Tandem configurations. *Software Engineering, IEEE Transactions on* **SE-12**(3) 450–461.

Porteus, Evan L. 2002. *Foundations of Stochastic Inventory Theory*. Stanford University Press.

Puterman, M.L. 1994. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, Inc.

Roger, JH. 1977. A significance test for cyclic trends in incidence data. *Biometrika* **64**(1) 152–155.

Roubos, D, S Bhulai. 2012. Approximate dynamic programming techniques for skill-based routing in call centers. *Probability in the Engineering and Informational Sciences* **26**(4) 581–591.

Roubos, Dennis, Sandjai Bhulai. 2010. Approximate dynamic programming techniques for the control of time-varying queuing systems applied to call centers with abandonments and retrials. *Probability in the Engineering and Informational Sciences* **24**(1) 27.

Ruth, R Jean. 1981. A mixed integer programming model for regional planning of a hospital inpatient service. *Management Science* **27**(5) 521–533.

Santibáñez, Pablo, Georgia Bekiou, Kenneth Yip. 2009. Fraser health uses mathematical programming to plan its inpatient hospital network. *Interfaces* **39**(3) 196–208.

Sauré, A., J. Patrick, S. Tyldesley, M.L. Puterman. 2012. Dynamic multi-appointment patient scheduling for radiation therapy. *European Journal of Operational Research* **223**(2) 573–584.

Smith, David R, Ward Whitt. 1981. Resource sharing for efficiency in traffic systems. *Bell System Technical Journal* **60**(1) 39–55.

Smith-Daniels, Vicki L, Sharon B Schweikhart, Dwight E Smith-Daniels. 1988. Capacity management in health care services: Review and future research directions. *Decision Sciences* **19**(4) 889–919.

Stroke Unit Trialists' Collaboration. 2007. Organised inpatient (stroke unit) care for stroke. *Cochrane Database Syst Rev* **4**(4).

Suri, Rajan, Gregory W Diehl. 1984. A new'building block'for performance evaluation of queueing networks with finite buffers. *ACM SIGMETRICS Performance Evaluation Review*. ACM, 134–142.

Takahashi, Yutaka, Hideo Miyahara, Toshiharu Hasegawa. 1980. An approximation method for open restricted queueing networks. *Operations Research* **28**(3-part-i) 594–602.

Thompson, Steven, Manuel Nunez, Robert Garfinkel, Matthew D Dean. 2009. Or practiceefficient short-term allocation and reallocation of patients to floors of a hospital

during demand surges. *Operations research* **57**(2) 261–273.

Tiwari, Vikram, H Sebastian Heese. 2009. Specialization and competition in healthcare delivery networks. *Health care management science* **12**(3) 306–324.

Utley, Martin, Steve Gallivan, Katie Davis, Patricia Daniel, Paula Reeves, Jennifer Worrall. 2003. Estimating bed requirements for an intermediate care facility. *European journal of operational research* **150**(1) 92–100.

Van Dijk, Nico, Erik van der Sluis. 2009. Pooling is not the answer. *European Journal of Operational Research* **197**(1) 415–421.

Vanberkel, Peter T, Richard J Boucherie, Erwin W Hans, Johann L Hurink, Nelly Litvak. 2012. Efficiency evaluation for pooling resources in health care. *OR spectrum* **34**(2) 371–390.

Vassilacopoulos, G. 1985. A simulation model for bed allocation to hospital inpatient departments. *Simulation* **45**(5) 233–241.

Weiss, Elliott N, Morris A Cohen, John C Hershey. 1982. An iterative estimation and validation procedure for specification of semi-markov models with application to hospital patient flow. *Operations Research* **30**(6) 1082–1104.

Weiss, Elliott N, John O McClain. 1987. Administrative days in acute care facilities: A queueing-analytic approach. *Operations Research* **35**(1) 35–44.

Whitt, Ward. 1993. Approximations for the gi/g/m queue. *Production and Operations Management* **2**(2) 114–161.

World Health Organization. 2006. *Neurological disorders: public health challenges*. World Health Organization.

Xie, Jipan, Eric Q Wu, Zhi-Jie Zheng, Janet B Croft, Kurt J Greenlund, George A Mensah, Darwin R Labarthe. 2006. Impact of stroke on health-related quality of life in the

noninstitutionalized population in the united states. *Stroke* **37**(10) 2567–2572.

Yonek, Joshi, S Hines, M Joshi. 2010. *A guide to achieving high performance in multi-hospital health systems*. Health Research & Educational Trust.

Yoon, Seunghwan, Mark E Lewis. 2004. Optimal pricing and admission control in a queueing system with periodically varying parameters. *Queueing Systems* **47**(3) 177–199.