A NEW APPROACH TO DYNAMIC ALLOCATION OF HOSPITAL ELECTIVE ADMISSIONS USING MARKOV DECISION PROCESSES

by

Amin Khoshkenar

A Thesis Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Industrial Engineering

Koç University

August 14, 2014

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Amin Khoshkenar

and have found that it is complete and satisfactory in all respects,

and that any and all revisions required by the final

examining committee have been made.

Committee Members:

____

Associate Professor Lerzan Örmeci

____

Assistant Professor Onur Kaya

____

Assistant Professor Evrim Didem Güneş

Date: ______________________

To my family

ABSTRACT

Uncontrolled patient flow and variability in hospital occupancy cause treatment delays, emergency blockages, cancellations, diversions and patient elopement. Efficient bed management through controlling inpatient admissions plays a significant role in decreasing waiting time, maintaining lower system occupancy and improving service delivery. Currently, for the sake of tractability, most researchers focus on a single ward, but unfortunately, in highly utilized hospitals, transferred patients from other wards make the admission gateway undesirable or infeasible both for patients and hospital managers. In this research, to obtain a better result, we allow two hospital wards to interact with each other and transfer their patients to the other's ward while there is no free bed in their own ward. We first present an optimal admission policy using a Markov decision process model, and then show some monotonicity results for optimal policies. In addition, we develop a set of sufficient conditions, which ensure that a patient type should be admitted or not, regardless of the system congestion level. Using the Markov decision process framework on inpatient admission, we show that leveraging controls to stabilize patient flow through the hospital wards can mitigate variability in hospital occupancy. This work is based on an actual private hospital concerned with resource allocation decisions and patient satisfaction improvements.

Key words: health care, hospital inpatient services, Markov decision processes, hospital admission control, bed occupancy, control of queuing systems.

ACKNOWLEDGMENTS

I would like to express my gratitude to my supervisor Lerzan Örmeci for the useful comments, remarks and engagement through the learning process of this master thesis. Furthermore, I would like to thank the rest of my thesis committee: Dr. Onur Kaya and Dr. Evrim Didem Güneş, for their insightful comments. Also, I like to thank the participants in my presentation, who have willingly shared their precious time. My gratitude also goes to the faculty members of Industrial Engineering Department at Koç University, there are not enough words to describe your excellent works. I would like to thank my parents and friends, who have supported me throughout entire process, both by keeping me harmonious and helping me putting pieces together. I will be grateful forever for your love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

In recent years, the need for managing hospital beds and other healthcare resources has increased considerably. Demand for medical care with high quality has been rising for years. Increasing numbers of admissions have put pressure on hospitals, and it can result in canceled operations, longer waiting lists and lower quality of care. Thanks to technological improvements in hospital services, lengths of stay have been decreasing and new ways of treating patients faster and in different settings are developed. The management of admission can be more effective in ensuring that beds are available for admitted patients, the quality of care is high and beds are used efficiently.

Admission to a hospital may be planned (elective) or urgency (emergency). Elective admissions are those who enter hospital through a medical consultant, a visit to the hospital outpatient department or a planned transfer from another hospital. Almost 40% of hospital admissions involve elective patients who have been on a waiting list, mostly for a surgical operation. While patients are waiting, they may be in pain and discomfort that interferes with their normal condition. Moreover, uncertainty can cause additional worry and in some cases, the patient's condition may deteriorate.

1.1 Variation and unpredictability

Unfortunately, the rising number of emergency admissions is only one of the sources of difficulty in controlling elective admissions. It is more problematic when there is variation in the number of patients seeking admission over time, and the unpredictability of this variation. This makes the management of elective admissions extremely complex and that is why admission scheduling and resource allocation in hospitals and medical centers is not an easy task. Admissions vary according to the time of day, or day to day with seasonal fluctuations (see Figure 5.1). For example, scarcity of beds is higher in winter than in summer, due to increased emergency admissions and longer lengths of stay.

The differing patterns of emergency and elective admissions offer different methods for managing hospital resources. If elective admissions can be controlled in a way that they are higher when emergency admissions are likely to be lower and vice versa, then the overall demand will be leveled. Managing the scheduled admissions is the first step in improving the admissions process.

1.2 Bed management

This work aim to evaluate inpatient admissions and bed management in a private hospital in Istanbul, Turkey. In this case, hospitals beds are considered as the main limited resource in the process of examination and treatment. Hospital managers are confronted with challenging questions such as: Are there available beds in the ward assigned to a department? What about other wards?

One of the most significant global healthcare issues relates to maximizing the efficiency of patient flow, while improving healthcare outcomes. We attempt to ensure bed availability for emergency patients who cannot wait by managing elective patients with respect to the emergency demand and allocating patients to their special or most appropriate unit. It requires to build a model, which enables us to have a smooth, safe, timely flow of patients through the hospital system and maximizing bed capacity without compromising patients' health. According to Figure 1.1, the right balance of capacity and process management achieves good patient outcomes (NHS, 1992).

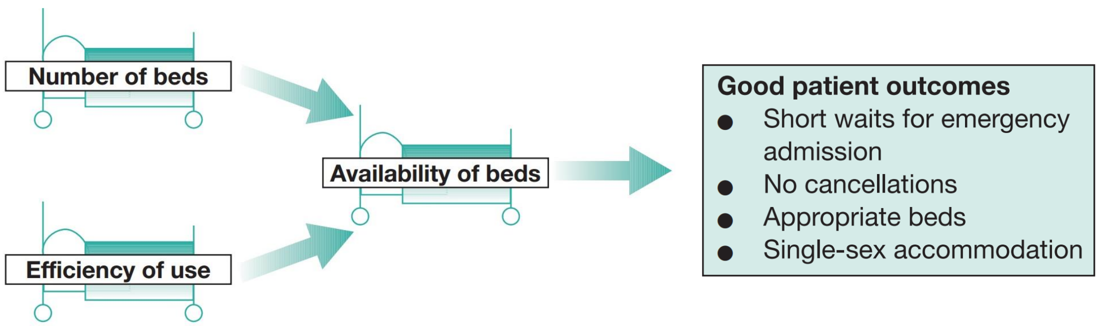

Figure 1.1: Effective bed management

In ward capacity allocation, most hospitals have to provide both for emergency admissions and elective admissions. Demand for emergency admission varies, so some spare capacity has to be available to meet the peaks in demand. Otherwise, on busy days, emergency admissions may displace elective admissions, leading to cancellations, disruption and inconvenience for patients. Patients cannot always be placed in the first available bed, or transferred to another ward. They must be directed to a ward, which matches their needs. Therefore, efficient managing of elective patients can make the situation better for all patients.

This research attempts to manage the hospital beds by optimizing patient admissions. Here, we allow two hospital wards to interact with each other. We present an optimal admission policy using a Markov decision process model and develop a set of sufficient conditions, which ensure that a patient type should be admitted or not, regardless of the system congestion level. Chapter 2 explores literature on health services, which uses the techniques like simulation, mathematical modeling, stochastic modeling and dynamic programming. Chapter 3 introduces the main model. Chapter 4 focuses on the structure of an optimal policy and shows certain properties of the model. Chapter 5 implements the model using a dataset from a Turkish private hospital and discusses the numerical results. Finally, chapter 6 concludes with the main findings and reviews the important contributions of the work.

Chapter 2

LITERATURE REVIEW

In recent years, the importance of healthcare systems has increased dramatically. The need for high customer satisfaction, improved healthcare quality and cost effective policies have attracted a considerable amount of researchers to work on this area. According to a survey conducted in 2004 on U.S. civilian non-institutionalized population, the inpatient stay in a hospital accounts for approximately one-third of all medical expenses, which is an important source of cost (Machlin and Carper, 2007).

2.1 Emergency Department

Hospital emergency departments (ED) play an important role in health care system, accounting for a increasing proportion of hospital admissions and serving as an advanced diagnostic center for primary care physicians. Although it is known as the most expensive place to obtain medical care, emergency rooms remain an important safety net for most people and may play a role in slowing the growth of health care costs. Some patients in ED need an inpatient care. However, because of a situation called access block, they are unable to access an appropriate care in a shorter period of time. Such situation results ED overcrowding (Fatovich et al., 2005). ED overcrowding threatens patient safety and originates from health care system failure and needs a multidisciplinary system approach to be less intense (Trzeciak and Rivers, 2003).

According to a review conducted by Hoot and Aronsky (2008) on articles related to ED crowding, commonly studied causes of crowding includes non-urgent visits, frequent-flyer patients, influenza season, inadequate staffing, inpatient boarding, and hospital bed shortages. Also, they find patient mortality, transport delays, treatment delays, ambulance diversion, patient elopement, and financial factors as important effects of crowding. An observational study was conducted by Forster et al. (2003) to identify the effect of hospital occupancy on ED length of stay. They conclude that increased hospital occupancy is strongly related with ED length of stay (LOS). ED overcrowding can also increase total admission LOS (Han et al., 2007).

Bed management plays an important role in improving service delivery, decreasing waiting time and maintaining lower bed occupancy (Proudlove et al., 2003). According to Sprivulis et al. (2006) analysis, there is a direct connection between efficient inpatient flow and appropriate bed usage. They also identify a linear relationship between the ED overcrowding and patient mortality rate. Based in an Erlang loss model, Bekker and Koeleman (2011) analyze the impact of variability in admissions and LOS on variability in bed occupancy. They apply a time-dependent analysis to determine the optimal number of admissions per day in such a way that average daily bed occupancy is obtained.

2.2 Simulation

Discrete-event simulation is a powerful tool in system analysis, which can be used to assess the efficiency of health care delivery systems, to explore possible relations among system variables, to design a new patient flow and to forecast the impact of some changes in the system in order to improve system performance and patient flow. New developments in simulation software technology have created numerous opportunities, including more sophisticated and valid analyses. Simulation models allow the end user (namely, hospital administrators or clinic managers) to assess the efficiency of the existing health care delivery systems, to ask 'what if?' questions, and to design new systems. In recent years, the application of discrete-event simulation in health care has become increasingly wide spread. Here, we review some of the articles that use simulation to analyze the patient flow.

For a general review on the use of discrete-event system simulation in healthcare facilities we refer to Jun et al. (1999). Their work shows that a large number of papers focus on the areas of resource allocation and patient flow. The use of computer simulation in the development of hospital systems, has a major effect on the cost and quality of operations. Harrison et al. (2005) use a multi-stage stochastic simulation model to find the daily occupancy distribution, to determine the range of normal fluctuations, and to investigate the trade-offs between frequency of overflows and the percent occupancy for bed allocations. They find seasonal variations and variable discharge rates to be significant factors in controlling overflow.

VanBerkel and Blake (2007) use a discrete event system simulation model to help the capacity planning of a division and to evaluate its performance. The model shows that long waiting times mainly depend on bed management and resource allocation. In another study using simulation methodology, Harper (2002) investigates how to manage the available hospital resources. His model is proved to be very helpful in the planning and management of hospital beds, operating theatres and workforce needs.

Harper and Shahani (2002) present a simulation model to evaluate the consequences of planning policies. Based on bed occupancies and refused admission rates, their model was able to effectively allocate the available beds. Lowery (1996) constructs a simulation model of patient flow in a hospital. The model is able to evaluate the effects of different scheduling systems, to determine appropriate allocations of beds and to provide a mechanism for making the hospital planning process.

2.3 Mathematical Modeling

Blakea and Carterb (2002) use linear programming for allocating resources in hospitals. The model was able to make decisions regarding the number, type and prices of services delivered by hospitals. Arenas et al. (2002) also construct a mathematical model to study the possibility of decreasing the LOS on the waiting list of a hospital. Belien and Demeulemeester (2007) use a mixed integer programming based heuristics and a meta-heuristic model to minimize the expected total bed shortage in a hospital. Their model is able to predict performance measures such as the daily expected bed occupancy, the variance on this occupancy, the expected bed shortage and the probability of a shortage on each day.

Roland et al. (2006) propose a mathematical model that takes into account resources and generates optimized plans. The objective of the model is to minimize the cost of opening new rooms and the overtime cost. They also develop a solution approach based on genetic algorithms. Vissers et al. (2005) develop a mixed integer linear model to evaluate scenarios for surgery planning in both tactical and strategic levels. The model is able to produce results in terms of resource utilizations and performance score.

2.4 Stochastic Models

Generally, many health service activities are best viewed as stochastic rather than deterministic processes. Basically, the reasons lie in the variability associated with the processes of illness and recovery. Consequently, stochastic modeling approach has predominated in, for example, studies of the admission, discharge and utilization of hospital inpatient facilities. Stochastic nature is the inherent part of a service system. Uncertainty in arrival and LOS is very common and stochastic models attempt to incorporate it. Mainly, these studies are based on some popular techniques like stochastic programming, queuing networks and Markovian models.

A queuing model was developed by Gorunescu et al. (2002a) to study the movement of patients through a hospital department, optimizing resources and controlling the bed occupancy. The model was able to optimally determine the number of beds in each ward given an acceptable level of patient rejection. Kao and Tung (1981) use a queuing approach to approximate the patient population dynamics with admission rates provided by forecasts. The model aims to minimize expected total average overflow and is validated by an empirical data set. The simplicity of the model and usefulness of the results make it a suitable approach to manage hospital wards.

Using queuing theory and data analysis, McManus et al. (2004) construct a model to evaluate the performance of a unit and manage its resources. Their model demonstrates that the required capacity of an intensive care unit (ICU) is significantly overestimated by managers. Moreover, the model is able to give a proper feedback whenever a sudden change occurs in the system. N. Koizumi (2005) use a queuing network with blocking to analyze the congestion in hospital facilities. They compare their mathematical model with a simulation model to find out the most critical facilities, in the sense that their shortage in the hospital may create significant problems in resource utilization.

de Bruin et al. (2007) explore how OR techniques have been applied in modeling the emergency cardiac inpatient flow. They show refused admissions are primarily caused by unavailability of beds in the hospital. Gorunescu et al. (2002b) propose an integrated model using queuing theory and computational techniques to improve the performance of bed allocation in such a way that the total cost is minimized. Generally, it enables decision makers to introduce different policies and to assess the benefit of providing extra beds to minimize the number of rejections when the demand increases.

2.5 Dynamic Programming

Markov decision processes (MDPs) are widely used in many industrial and manufacturing applications such as logistics, finance and inventory control (Puterman, 1994). We review the papers in this area that are close to ours.

Ross (1995) introduces basic stochastic knapsack which consists of s servers with no waiting room to give service to k classes. Each class is distinguished by its size, its arrival rate and its mean service time.

Hajek (1984) presents a cost-based MDP model to describe a service system, where two stations are interacting with each other. They proved the existence of optimal controls which have a switch curve structure. Ghoneim and S. Stidham (1985) consider two queues in series with input to each queue, which can be controlled by accepting or rejecting arriving customers. They maximized the discounted or average expected net benefit over a finite or infinite horizon, and proved that their model is monotonic in some senses.

Miller (1969) considers a reward-based multi-server queuing system, which serves different customer classes. For a single server with 2 customer types, he showed it is optimal to accept class-1 customer whenever there is an idle server and to admit class-2 customer if and only if the total number of customers in the system is less than a certain threshold. Lippman (1975) studies the problem of controlling M/M/c queuing systems and obtained the optimal policy for a system, which serves different customer classes.

Altman et al. (1998) consider call admission control of multiple classes to a group of servers without waiting room and presented some structural results. Ormeci et al. (2001) study the problem of dynamic admission control system in a two class loss Markovian queueing system by considering different service rates for each customer type. They determined an optimal admission policy and developed a set of sufficient conditions which ensure that a customer class should be admitted or not. Savin et al. (2005) construct a queueing control model to study the allocation of capacity in a system in which rental equipment is accessed by two classes of customers. They use dynamic programming to characterize properties of optimal capacity allocation policies and find preferred customer classes under some specific conditions.

Ormeci (2004) investigates the admission control of a call center with two classes of calls and three stations, one dedicated for each class, and one shared station. She shows that serving a call in its dedicated station, whenever possible, is optimal and proves the existence of optimal monotone thresholds under certain conditions. Helm et al. (2011) use MDP to design an admission control system for cost effectiveness.

Carrizosa et al. (1998) consider a loss model, where different types of customers arrive according to a Poisson process with different rates (heterogeneous arrivals). They assumed duration of service for each customer is a random variable with a finite mean, and presented an optimal static control policy for acceptance/rejection of different types of customers. Ku and Jordan (1997) analyze access control policies for two stations in tandem each with no waiting room and parallel servers. They also found the optimal policy is monotonically decreasing threshold and proved monotonic variation of the threshold for most system parameters.

Lewis et al. (1999) present a finite capacity reward system to describe the structure of stationary, deterministic and optimal policies.Hordijk and Koole (1992) use a dynamic model to describe parallel queues in which each queue has its own single server, and service times are exponential with non-identical parameters. They also showed some optimality results for their cost-based model.

Green et al. (2006) formulate a finite-horizon dynamic programming to allocate the impatient into a diagnostic medical facility. Gerchak et al. (1996) use a stochastic dynamic programming for elective surgery scheduling regarding uncertain capacity of operating room. Patrick et al. (2008) propose a Markov Decision Process model to dynamically allocate available capacity to patients with different priorities. Because the state space was too large, they used approximate dynamic programming approach to solve the model. Dobson et al. (2011) formulate a MDP to analyze the effect of reserving slots for urgent patients in a healthcare setting.

Chapter 3

MARKOV DECISION PROCESS MODEL

3.1 Markov Decision Processes

Markov Decision Processes (MDPs) have attracted the attention of many researchers because they are important both from the practical and the intellectual points of view. MDPs or stochastic dynamic programs provide a general framework for modeling dynamic systems under uncertainty. These models provide tools for the solution of important healthcare problems. In particular many business and engineering applications use MDP models. Analysis of various problems arising in MDPs leads to a large variety of interesting mathematical and computational problems.

One of the popular applications of MDPs is to control the patient flow in hospitals and medical centers. The purpose of controlling patient admissions is to promote a more efcient utilization of hospital resources. In some hospital admission systems, decisions must be made sequentially in an uncertain environment. Hospital manager must consider the currently available resources as well as the potential future. Often decisions are to be taken in a dynamic environment. MDP is an appropriate technique that is useful in a class of problems involving complex, stochastic and dynamic decisions for which it can reveal optimal solutions. The goal of a MDP model is to provide an optimal policy that optimizes a particular criterion such as maximizing total discounted reward or minimizing the total discounted cost.

MDPs link previous, current and future decisions through the proper definition of system states defined as variables that contain the relevant information for making decisions. The decision model evolves in the following manner: The state of the system, say capacity, is observed (or partially observed), an action is taken, a cost is incurred (or a reward is received) and the system move into a new state according to a known probability distribution, which depends on the state and the action. In general, a decision rule can depend not only on the current state of the system, but also on all previous states and actions. However, in a MDP model the future transitions are independent of the past, which is the standard assumption of a Markov process. The key ingredients of this model are the following:

- a set of decision times (epochs),
- a set of system states (state space),
- a set of available actions,
- the state- and action-dependent rewards or costs,
- the state- and action-dependent transition probabilities on the state space.

Given such a model, we would like to know how the admission manager should act so as to minimize the costs, possibly subject to some constraints on the allowed states of the system. To this end, we will be interested in finding optimal decision rules, which specify the action be chosen at a particular epoch in each state.

MDPs also have drawbacks. As the size of the problem increases, MDPs become harder to solve exactly. However, many techniques for finding approximate solutions to MDPs exist. Perhaps the largest problem to the broader application of MDPs is the data. Obtaining quality medical data is very difficult and expensive (Schaefer et al., 2004).

3.1.1 Finite-horizon MDPs

Following the notation of Puterman (1994), the basic model of a finite horizon, discrete-time MDP is defined by $(S, A, p_t(.|.,.), c_t(.,.), N)$, where S is the set of state $s \in S$, A is the set of all feasible actions or decisions and A_s are those actions available at state s. The system progresses to state s' from state s when action $a \in A_s$ is chosen at decision epoch t, $(t = 1, 2, ..., N)$, with a known transition probability $p_t(s'|s, a)$. When action $a \in A_s$ is chosen in state s at decision epoch t, a cost $c_t(s, a)$ is incurred. We define a policy $\pi = \{d_1(s), d_2(s), ..., d_{N-1}(s)\}$ as a sequence of decision rules, where a decision rule is a mapping from states to actions, so that $d_t(s) \in A_s$. The application of a policy π induces a probability distribution over the states at various stages, where the state of the system after t transitions is X_t and the action chosen, Y_t, is a function of this state. The objective is to compute the policy that minimizes a given criterion in expectation.

There are mainly three commonly used method to solve a MDP model: total expected average cost, total discounted expected cost and linear programming. Here we focus on the total discounted expected cost, which is defined as follows:

$$V_N^\pi(s) = E_s^\pi \left\{ \sum_{t=1}^{N-1} \beta^{t-1} c_t(X_t, Y_t) + \beta^{N-1} c_N(X_N) \right\},$$

where $0 \leq \beta < 1$ is a discount factor.

The separability of the MDP decisions and costs can be exploited to decompose this N-period problem into a sequence of $N - 1$ single-stage problems, by recursively solving backward from stage $N-1$ to 1:

$$V_N(s) = c_N(s) \ \mathit{for\ every}\ s \in S,\ \mathit{and}$$

$$V_t(s) = \min_{a \in A_s} \left\{ c_t(s,a) + \sum_{s' \in S} \beta p(s'|s,a) V_{t+1}(s') \right\},$$

for $t = N-1, ..., 1$ and $s \in S$. Here $V_t(s)$ is the total discounted expected cost of the $N-t$ stage problem beginning in state s at stage t or as a single-stage problem with terminal costs $V_{t+1}(s')$, which are known at the time $V_t(s)$ is computed. This is the true computational benefit of MDPs, the ability to reduce a problem into manageable subproblems and still attain the optimal solution. The optimal policy is defined to be the sequence of decision rules, mapping states to the actions that minimize the above recursion, i.e.

$$d_t(s) = arg \min_{a \in A_s} \left\{ c_t(s,a) + \sum_{s' \in S} \beta p(s'|s,a) V_{t+1}(s') \right\},$$

for $t = N-1, ..., 1$ and $s \in S$. In the above solution and model we assumed that the decision horizon was a finite N. Often there is no defined horizon or the number of stages is so high that it can be approximated by an infinite horizon.

3.1.2 Infinite-horizon MDPs

Infinite-horizon models require an infinite amount of data. Therefore, it is typically assumed that data are time-homogeneous or changing so slowly that homogeneity is a reasonable assumption. As a result, the state of an infinite horizon MDP must be carefully defined to ensure that the system transitions are stationary. If the data are naturally time-dependent, the time-homogeneity assumption can be satisfied by properly augmenting the state definition with the time at which a system transition occurs. The presence of stationary system transitions allows for the use of several elegant solution techniques and easily characterized optimal policies. We replace the above finite-horizon criteria with an infinite-horizon variant by taking the limit of each measure as N goes to infinity.

One of the key insights into infinite-horizon MDPs is that as a result of the assumptions of an infinite horizon, time-homogeneity, and Markov property, under a stationary policy π, i.e. $d_t(s) = d(s)$ for all $s \in S$ and $t = 1, 2, ...$, the expected reward vector is also stationary.

$$V_t^{\pi}(s) = V^{\pi}(s) \ for \ t = 1, 2, ..., and \ s \in S,$$

and $V^{\pi}(s)$ is the unique solution to the set of equations:

$$V^{\pi}(s) = c(s, d(s)) + \sum_{s' \in S} \beta p(s'|s, d(s)) V^{\pi}(s') \ for \ s \in S.$$

It is well known that a stationary policy is optimal for these MDPs (Puterman, 1994). In addition, the optimal vector V^* is the solution of the following equations, known as Bellmans equations:

$$V^*(s) = min_{(a \in A_s)} \left\{ c(s, a) + \sum_{s' \in s} \beta p(s'|s, a) V^*(s') \right\} \ for \ s \in S.$$

Given any initial bounded vector V_0, it can be shown that the following sequence converges to a solution of Bellmans equations:

$$V_k(s) = min_{(a \in A_s)} \left\{ c(s, a) + \sum_{s' \in s} \beta p(s'|s, a) V_{k-1}^*(s') \right\} \ for \ s \in S \ and \ k = 1, 2, ...$$

However, this solution procedure, known as value iteration, may require an infinite number of iterations. As a result, another technique, policy iteration, is typically used to search over the finite space of policies.

3.2 A Markov Decision Process Model for Hospital Bed Management

Beside the modeling part, we provided a comprehensive data analysis to analyze the admission and length of stay (LOS) processes. Outputs are based on the computerized records of all admissions to a private hospital over the years 2011 and 2012. We covered the following departments:

- Cardiovascular Surgery,
- ENT,
- General Surgery,
- Neurosurgery,
- Obstetrics and Gynecology,
- Orthopedics,
- Pediatrics,
- Plastic surgery,
- Transplantation,
- Urology.

In this part we attempt to construct a model to control the patient flow in these departments. There are mainly two source of arrivals to each department; arrivals from outside and transferred patients from other wards. Building a comprehensive model to cover all of the departments and control their admission process is not an easy task; however, it is possible to construct some useful models. In this work, we grouped all of the hospital wards into two parts. One group only contains a single hospital ward, which we want to focus on, and the other one consists of the remaining wards. Each group works as a service station in the model.

The model mainly contains two hospital wards. One of them is a particular ward, and the other one is a dummy ward that includes all the other potential wards that a patient can transfer. Figure 3.1 shows that patients are separated into two groups. Type 1 patients are served in ward 1, and the other group in Ward 2. When an elective type i patient arrives, he/she would be either rejected with a cost of c_i , or would be accepted in his/her own ward, or would be transferred to the other ward with a cost of t_i (per time unit when he/she stays there) if there is no free bed in his/her own ward. Also, when an emergency patient arrives, with probability p_i, he/she is characterized as type i patient, and directly accepted in his/her own ward, or would be transferred to the other ward with a cost of t_i if there is no free bed in his/her own ward. Note that $p_1 + p_2 = 1$. We always want to accept the emergency patients, so if there is no free bed in the hospital upon a emergency arrival, it will cost M (large number).

Emergency and elective type i patients arrive to the hospital according to a Poisson process with rates λ_e and λ_i respectively. Once he/she is admitted to his/her own ward, he/she will stay there according to an exponential distribution with rate μ_i. If he/she is admitted, but transferred to the other ward, then he/she will stay in hospital according to an exponential distribution with rate $\frac{\mu_i}{1+\gamma}$.

Let $X^\pi(t) = (x_1^\pi(t), x_2^\pi(t))$ denotes the state of the system at time t under policy π, where $x_i^\pi(t) \in \mathbb{Z}^+$ is the number of type i patients in the hospital for $i = 1, 2$. We will use $X = (x_1, x_2)$ instead of $X^\pi(t) = (x_1^\pi(t), x_2^\pi(t))$. Let B_i

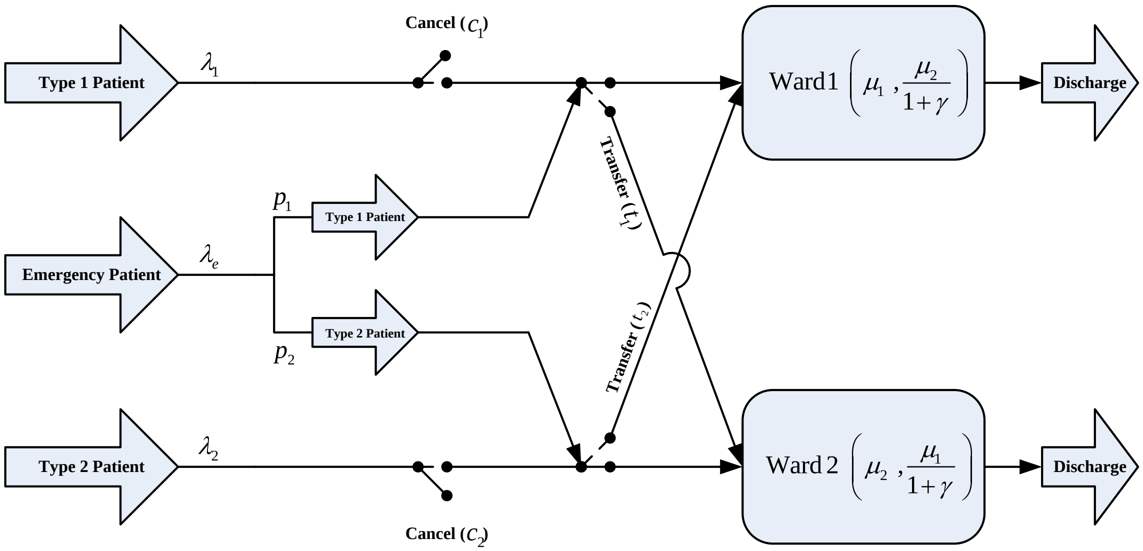

Figure 3.1: Admission control system

denote the active capacity of the ward assigned to type i patients. The state space is $\Psi = \{X : x_1 + x_2 \leq B_1 + B_2\}$. Our goal is to analyze the structure of the optimal policy for this model and to gain insights from this analysis into hospital management practice.

3.3 Assumptions and definitions

Generally, the arrival rates vary in a non-stationary manner by the time of day and day of week. However, it has been widely accepted that arrival process, especially unscheduled arrivals, behaves according to a Poisson process (Harrison et al., 2005; Young, 1965; Bruin et al., 2010). In our case, for each department the admission pattern is compared with the Poisson distribution and for most of the departments many arrivals have been shown to be well approximated by a Poisson process (See section 5.1). Additionally, the following assumptions are considered.

- Canceling an operation is more costly than admitting by transferring, so $c_i \geq t_i$; for $i = 1, 2$. Note that t_i is incurred per patient per time unit.

- Type 1 patients are more serious than the type 2 patients, so that type 2 patients leave the ward faster than type 1 patients, so we assume that $\mu_2 \geq \mu_1$; for $i = 1, 2$,

- The previous assumption even when a type 2 patient is transferred holds. Therefore, $\frac{\mu_2}{1+\gamma} \geq \mu_1$,

- Patients are moved to their own ward when there is a free space there.

- We set $[a - b]^+ = max(a - b, 0)$,

- e_i represents i^{th} unit vector,

- i^c denotes $\{1, 2\} - i$.

3.4 Markov Decision Process Formulation

We develop a discounted discrete time MDP model over a finite horizon. We apply uniformization (Lippman, 1975). As we assume, without loss of generality, $\mu_2 \geq \mu_1$ and $\frac{\mu_2}{1+\gamma} \geq \mu_1$ (patient type 1 stays more than type 2). Then, the maximum possible rate out of any state is $\lambda_1 + \lambda_2 + \lambda_e + (\frac{1}{1+\gamma}B_1 + B_2)\mu_2$. However, certain parts of the analysis requires a symmetric model. Accordingly, the uniformization factor is defined as $\varphi = \lambda_1 + \lambda_2 + \lambda_e + (\mu_1 + \mu_2)(B_1 + B_2)$. Since the time between transitions is always exponentially distributed and the uniformization factor is finite, we can use uniformization and normalization to build a discrete time equivalent of the original system. We assume, using the appropriate time scale, $\lambda_1 + \lambda_2 + \lambda_e + (\mu_1 + \mu_2)(B_1 + B_2) = 1$, so that the system will be observed at exponentially distributed intervals with mean 1. There will be an arrival with probability $\lambda_1 + \lambda_2 + \lambda_e$ and a potential service completion with probability $(\mu_1 + \mu_2)(B_1 + B_2)$. Let β be the continuous time discount factor. Now we can formulate the finite-horizon optimal expected discounted cost recursive optimality equation as:

$$
\begin{aligned}
V_{n+1,\beta}(X) = \sum_{i=1}^{2} t_i[x_i - B_i]^+ + \beta\Bigg\{ & \sum_{i=1}^{2} \lambda_i \min\left\{V_{n,\beta}(X) + \frac{c_i}{\beta}, V_{n,\beta}(X+e_i)\right\} \\
& + \sum_{i=1}^{2} p_i \lambda_e V_{n,\beta}(X+e_i) \\
& + \sum_{i=1}^{2} \min(x_i, B_i)\mu_i V_{n,\beta}(X-e_i) + \sum_{i=1}^{2}[x_i - B_i]^+ \frac{\mu_i}{(1+\gamma)} V_{n,\beta}(X-e_i) \\
& + \left(\sum_{i=1}^{2}(B_i + B_{i^c})\mu_i - \sum_{i=1}^{2}\left(\min(x_i, B_i)\mu_i + [x_i - B_i]^+ \frac{\mu_i}{(1+\gamma)}\right)\right) V_{n,\beta}(X)\Bigg\},
\end{aligned}
\tag{3.1}
$$

where $V_{n,\beta}(X)$ represents the optimal cost of the n-period β-discounted problem starting in state $X = (x_1, x_2)$. According to Eq. (3.1), the first term of the value function is the one period instantaneous cost. The system incurs a cost of t_i when a type i patient stays in the other ward. The second term represents the arrival of elective patients. When this event occurs, the controller either admits the incoming patient to the hospital and assigns a bed upon arrival (term 2 of the minimization) or rejects them (term 1). The third term represents the arrival of emergency patients, which are always accepted if there is a free bed in the system. The fourth term represents the regular discharge of a patient from his/her own ward. The fifth term represents the discharge of a patient from the other ward. When a patient is discharged, number of occupied beds by type i patients decreases from x_i to $x_i - 1$. Finally, the last term corresponds to the fictitious events according to the uniformization factor φ.

3.5 Event-Based Dynamic Programming

Event-based dynamic programming (EBDP) deals with event operators, which can be seen as building blocks of the value function. Typically, we associate an operator with each basic event in the system, such as an arrival at a queue, a service completion, etc. EBDP focuses on the underlying properties of the value and cost functions, and allows us to study many models at the same time. According to Koole (1998), reformulation of problem using EBDP is suggested for convenience of analysis. Using EBDP we proceed as follows. For any function $f : \mathbb{N}_0 \to \mathbb{R}$, Let uniformization operators $T_{unif}[f_1, f_2, f_3, f_4, f_5](X)$ and $T_{cost}f(X)$ as:

$$T_{unif}[f_1, f_2, f_3, f_4, f_5](X) = \lambda_1\beta f_1(X) + \lambda_2\beta f_2(X) + \lambda_e\beta f_3(X) + \mu_1\beta f_4(X) + \mu_2\beta f_5(X), \tag{3.2}$$

$$T_{cost}f(X) = C(X) + f(X), \tag{3.3}$$

where $C(X)$ is defined as:

$$C(X) = \sum_{i=1}^{2} t_i[x_i - B_i]^+.$$

For the elective arrivals, we define operator $T_{AR_i}f(X)$ for $i = 1, 2$ as follow:

$$T_{AR_i}f(X) = min\{f(X) + \frac{c_i}{\beta}, f(X + e_i)\}, \tag{3.4}$$

and for emergency arrivals, we define operator $T_E f(X)$ as follow:

$$T_E f(X) = \begin{cases} p_1 f(X + e_1) + p_2 f(X + e_2) & if \quad x_1 + x_2 < B_1 + B_2, \\ M + f(X) & if \quad x_1 + x_2 = B_1 + B_2, \end{cases} \tag{3.5}$$

which M is a big number. T_{unif}, T_{cost}, T_{AR_i} and $T_E f(X)$ are four standard operators introduced in Koole (1998). Moreover, we want to define a discharge operator. According to (3.1), for discharge part we have:

$$\min(x_i, B_i)\mu_i V_{n,\beta}(X - e_i) + [x_i - B_i]^+ \frac{\mu_i}{(1+\gamma)} V_{n,\beta}(X - e_i)$$
$$+ \left((B_1 + B_2)\mu_i - \sum_{i=1}^{2} \left(\min(x_i, B_i)\mu_i + [x_i - B_i]^+ \frac{\mu_i}{(1+\gamma)} \right) \right) V_{n,\beta}(X).$$

By replacing $V_{n,\beta}$ with f, we have:

$$\min(x_i, B_i)\mu_i f(X - e_i) + [x_i - B_i]^+ \frac{\mu_i}{(1+\gamma)} f(X - e_i)$$
$$+ \left((B_1 + B_2)\mu_i - \left(\min(x_i, B_i)\mu_i + [x_i - B_i]^+ \frac{\mu_i}{(1+\gamma)} \right) \right) f(X).$$

For case $x_i \le B_i$, we have:

$$x_i\mu_i f(X - e_i) + ((B_1 + B_2)\mu_i - x_i\mu_i) f(X)$$
$$= \mu_i \underbrace{(x_i f(X - e_i) + ((B_1 + B_2) - x_i) f(X))}_{T_{D_i} f(X)}.$$

Similarly, for case $x_i > B_i$ we have:

$$B_i\mu_i f(X - e_i) + \frac{\mu_i(x_i - B_i)}{(1+\gamma)} f(X - e_i) + \left((B_1 + B_2)\mu_i - \left(B_i\mu_i + \frac{\mu_i(x_i - B_i)}{(1+\gamma)} \right) \right) f(X)$$
$$= \mu_i \underbrace{\left(\frac{\gamma B_i + x_i}{1+\gamma} f(X - e_i) + \left(\frac{(1+\gamma)B_{i^c} + B_i - x_i}{1+\gamma} \right) f(X)) \right)}_{T_{D_i} f(X)}.$$

Therefore, we define $T_{D_i} f(X)$ for all $X \in \Psi$ and $i = 1, 2$ as:

$$T_{D_i} f(X) = \begin{cases} \frac{\gamma B_i + x_i}{1+\gamma} f(X - e_i) + \left(\frac{(1+\gamma)B_{i^c} + B_i - x_i}{1+\gamma} \right) f(X), & if \quad x_i > B_i, \\ x_i f(X - e_i) + (B_1 + B_2 - x_i) f(X), & if \quad x_i \le B_i. \end{cases} \tag{3.6}$$

These Ts are called the *event operators*. Using these operators, we can rewrite the value function as:

$$V'_{n+1,\beta}(X) = T_{cost}.T_{unif}[T_{AR_1} V'_{n,\beta}, T_{AR_2} V'_{n,\beta}, T_E V'_{n,\beta}, T_{D_1} V'_{n,\beta}, T_{D_2} V'_{n,\beta}](X). \tag{3.7}$$

These operators can be used to study all kinds of monotonicity results.

Chapter 4

STRUCTURE OF AN OPTIMAL POLICY

The following sections derive the structure of an optimal policy using the n-period discounted model described in (3.1). We analyze the model to gain insights for a policy that balances the hospital's needs for operational efficacy.

4.1 Basic properties

We first show the monotonicity of the value function. We want to see how the value function changes by adding an additional patient to the system. Intuitively, by increasing the number of patients in the system, cost function V_n increases as well as shown in the next Theorem.

Theorem 4.1. *For all X, j and n, $V_n(X+e_j) \geq V_n(X)$.*

Proof. For $j=2$, $V_{0,\beta}(X) \equiv 0\ \forall X$, so it satisfies the inequality. Let f be a function with $f(X+e_j) \geq f(X)$. Now we show that $Tf(X+e_j) \geq Tf(X)$ for all of the operators defined in the last chapter. For operator T_{AR_i}, we consider the following two cases:

Case 1: $min\{f(X+e_2)+\frac{c_i}{\beta}, f(X+e_2+e_i)\} = f(X+e_2)+\frac{c_i}{\beta}$

$$\begin{aligned}
T_{AR_i}f(X) = min\{f(X)+\frac{c_i}{\beta}, f(X+e_i)\} &\leq f(X)+\frac{c_i}{\beta} \\
&\leq f(X+e_2)+\frac{c_i}{\beta} \\
&= min\{f(X+e_2)+\frac{c_i}{\beta}, f(X+e_2+e_i)\} \\
&= T_{AR_i}f(X+e_2)
\end{aligned}$$

$$\Rightarrow \qquad T_{AR_i}f(X) \leq T_{AR_i}f(X+e_2)$$

$min\{f(X)+\frac{c_i}{\beta}, f(X+e_i)\}$ is always less than or equal to $f(X)+\frac{c_i}{\beta}$. Since we assumed $f(X+e_j) \geq f(X)$, so for $j=2$, $f(X)+\frac{c_i}{\beta} \leq f(X+e_2)+\frac{c_i}{\beta}$. Finally, according to case 1 assumption, $f(X+e_2)+\frac{c_i}{\beta} = min\{f(X+e_2)+\frac{c_i}{\beta}, f(X+e_2+e_i)\}$, which is $T_{AR_i}f(X+e_2)$. Similarly, for Case 2 we have:

Case 2: $min\{f(X+e_2)+\frac{c_i}{\beta}, f(X+e_2+e_i)\} = f(X+e_2+e_i)$

$$\begin{aligned}
T_{AR_i}f(X) = min\{f(X)+\frac{c_i}{\beta}, f(X+e_i)\} &\leq f(X+e_i) \\
&\leq f(X+e_2+e_i) \\
&= min\{f(X+e_2)+\frac{c_i}{\beta}, f(X+e_2+e_i)\} \\
&= T_{AR_i}f(X+e_2)
\end{aligned}$$

$$\Rightarrow \qquad T_{AR_i}f(X) \leq T_{AR_i}f(X+e_2)$$

Hence, $T_{AR_i}f(X+e_2) \geq T_{AR_i}f(X)$; for $i=1,2$.

For operator T_{D_1}, we have:

Case 1: $x_1 > B_1, X \in \Psi$

$$\begin{aligned}
T_{D_1}f(X) &= \frac{\gamma B_1 + x_1}{1+\gamma} f(X-e_1) + \left(\frac{(1+\gamma)B_2 + B_1 - x_1}{1+\gamma}\right) f(X) \\
&\leq \frac{\gamma B_1 + x_1}{1+\gamma} f(X-e_1+e_2) + \left(\frac{(1+\gamma)B_2 + B_1 - x_1}{1+\gamma}\right) f(X+e_2) \\
&= T_{D1}f(X+e_2)
\end{aligned}$$

$$\Rightarrow \qquad T_{D_1}f(X) \leq T_{D1}f(X+e_2)$$

The first term is $T_{D_1}f(X)$ when $x_1 > B_1, X \in \Psi$. Since $f(X-e_1+e_2) \geq f(X-e_1)$, so $\frac{\gamma B_1+x_1}{1+\gamma}f(X-e_1+e_2) \geq \frac{\gamma B_1+x_1}{1+\gamma}f(X-e_1)$, and similarly $\left(\frac{(1+\gamma)B_2+B_1-x_1}{1+\gamma}\right) f(X+e_2) \geq \left(\frac{(1+\gamma)B_2+B_1-x_1}{1+\gamma}\right) f(X)$.

Case 2: $x_1 \le B_1, X \in \Psi$

$$\begin{aligned} T_{D_1}f(X) &= x_1 f(X - e_1) + (B_1 + B_2 - x_1)\, f(X) \\ &\le x_1 f(X - e_1 + e_2) + (B_1 + B_2 - x_1)\, f(X + e_2) \\ &= T_{D1} f(X + e_2) \\ &\quad \Rightarrow \qquad T_{D_1}f(X) \le T_{D1} f(X + e_2) \end{aligned}$$

Therefore, $T_{Di}f(X + e_2) \ge T_{Di}f(X)$, for $i = 1, 2$. For rest of the operators, we can simply prove the inequality. □

We also need to present the following theorem. It indicates that $V_n(X) - V_n(X + e_i)$ is bounded for $i = 1, 2$.

Theorem 4.2. *If $f(X) - f(X + e_i) \le M$, then $Tf(X) - Tf(X + e_i) \le M$ for $i = 1, 2$.*

Proof. We need to show this for all of the operators. For operator T_E, let's consider case $i = 1$. We have two possible cases:

Case 1: accept in $X + e_1$

$$T_E f(X) - T_E f(X + e_1) = p_1 f(X + e_1) + p_2 f(X + e_2) - p_1 f(X + 2e_1) - p_2 f(X + e_1 + e_2)$$

$$= p_1(f(X + e_1) - f(X + 2e_1)) + p_2(f(X + e_2) - f(X + e_1 + e_2))$$

$$\le p_1(f(X) - f(X + e_1)) + p_2(f(X) - f(X + e_1))$$

$$= f(X) - f(X + e_1) \le M.$$

Case 2: reject in $X + e_1$

$$T_E f(X) - T_E f(X + e_1) = p_1 f(X + e_1) + p_2 f(X + e_2) - M - f(X + e_1)$$

$$= p_2(f(X + e_2) - f(X + e_1)) - M \le M.$$

Similarly, we can prove for case $i = 2$. The same procedure can be used to prove for other operators as well. □

4.2 Supermodularity

In order to show that there exists an optimal policy that is of threshold type, we need to prove the following property, which is called *supermodularity.*

Theorem 4.3. *The value function $V_{n,\beta}(X)$ is supermodular, i.e.,*

$$V_{n,\beta}(X) - V_{n,\beta}(X+e_1) - V_{n,\beta}(X+e_2) + V_{n,\beta}(X+e_1+e_2) \geq 0.$$

Proof. In order to show the supermodularity for $V_{n,\beta}(X)$, it has to be shown that this property is satisfied by all of the operators. Since $V_{0,\beta}(X) \equiv 0$ for all X, $V_{0,\beta}$ is trivially supermodular.

Supermodularity of T_{unif}:

First assume f_1, f_2, f_3, f_4 and f_5 are suppermodular. We have:

$$T_{unif}[f_1,f_2,f_3,f_4,f_5](X) - T_{unif}[f_1,f_2,f_3,f_4,f_5](X+e_1) - T_{unif}[f_1,f_2,f_3,f_4,f_5](X+e_2)$$
$$+T_{unif}[f_1,f_2,f_3,f_4,f_5](X+e_1+e_2) \geq 0,$$

which can be written as follows:

$$\lambda_1\beta f_1(X)+\lambda_2\beta f_2(X)+\lambda_e\beta f_3(X)+\mu_1\beta f_4(X)+\mu_2\beta f_5(X)-\lambda_1\beta f_1(X+e_1)-\lambda_2\beta f_2(X+e_1)$$
$$-\lambda_e\beta f_3(X+e_1)-\mu_1\beta f_4(X+e_1)-\mu_2\beta f_5(X+e_1)-\lambda_1\beta f_1(X+e_2)-\lambda_2\beta f_2(X+e_2)$$
$$-\lambda_e\beta f_3(X+e_2)-\mu_1\beta f_4(X+e_2)-\mu_2\beta f_5(X+e_2)+\lambda_1\beta f_1(X+e_1+e_2)+\lambda_2\beta f_2(X+e_1+e_2)$$
$$+\lambda_e\beta f_3(X+e_1+e_2)+\mu_1\beta f_4(X+e_1+e_2)+\mu_2\beta f_5(X+e_1+e_2) \geq 0.$$

After rewriting, we have:

$$\lambda_1\left(f_1(X) - f_1(X+e_1) - f_1(X+e_2) + f_1(X+e_1+e_2)\right)$$
$$+\lambda_2\left(f_2(X) - f_2(X+e_1) - f_2(X+e_2) + f_2(X+e_1+e_2)\right)$$

$$+\lambda_e \left(f_3(X) - f_3(X+e_1) - f_3(X+e_2) + f_3(X+e_1+e_2)\right)$$

$$+\mu_1 \left(f_4(X) - f_4(X+e_1) - f_4(X+e_2) + f_4(X+e_1+e_2)\right)$$

$$+\mu_2 \left(f_5(X) - f_5(X+e_1) - f_5(X+e_2) + f_5(X+e_1+e_2)\right) \geq 0,$$

which holds true from supermodularity of f_1, f_2, f_3, f_4 and f_5.

Supermodularity of T_{cost}:

$$T_{cost}f(X) - T_{cost}f(X+e_1) - T_{cost}f(X+e_2) + T_{cost}f(X+e_1+e_2) \geq 0.$$

$$C(X)+f(X)-C(X+e_1)-f(X+e_1)-C(X+e_2)-f(X+e_2)+C(X+e_1+e_2)+f(X+e_1+e_2) \geq 0.$$

$$C(X)-C(X+e_1)-C(X+e_2)+C(X+e_1+e_2)+\underbrace{f(X) - f(X+e_1) - f(X+e_2) + f(X+e_1+e_2)}_{\text{this is positive because of supermodularity of } f.}$$

$$\geq 0,$$

so it remains to show:

$$C(X) - C(X+e_1) - C(X+e_2) + C(X+e_1+e_2) \geq 0.$$

The cost term for class-i patient can be written as follow:

$$t_i[x_i - B_i]^+ - t_i[x_i + 1 - B_i]^+ - t_i[x_i - B_i]^+ + t_i[x_i + 1 - B_i]^+ = 0,$$

which satisfies the inequality. Therefore T_{cost} is supermodular.

Supermodularity of T_{AR_i}:

For $i = 1$, each one of $T_{AR_1}f(X)$ and $T_{AR_1}f(X+e_1+e_2)$ can either accept or reject an incoming patient. Therefore, we have four possible cases, which are as follows:

Case 1: accept in X, accept in $X + e_1 + e_2$:

$$\begin{aligned} T_{AR_1} f(X + e_1) + T_{AR_1} f(X + e_2) &\leq f(X + 2e_1) + f(X + e_1 + e_2) \\ &\leq \underbrace{f(X + e_1)}_{\text{accept}} + \underbrace{f(X + 2e_1 + e_2)}_{\text{accept}} \\ &= T_{AR_1} f(X) + T_{AR_1} f(X + e_1 + e_2) \end{aligned}$$

The first inequality is true since accept is a feasible action in both states $X + e_1$ and $X + e_2$. The second inequality follows from the supermodularity of f, and since the Case 1 accepts in X, and $X + e_1 + e_2$, so $f(X + e_1) = T_{AR_1} f(X)$ and $f(X + 2e_1 + e_2) = T_{AR_1} f(X + e_1 + e_2)$.

Case 2: reject in X, reject in $X + e_1 + e_2$:

$$\begin{aligned} T_{AR_1} f(X + e_1) + T_{AR_1} f(X + e_2) &\leq f(X + e_1) + \frac{c_1}{\beta} + f(X + e_2) + \frac{c_1}{\beta} \\ &\leq f(X) + \frac{c_1}{\beta} + f(X + e_1 + e_2) + \frac{c_1}{\beta} \\ &= T_{AR_1} f(X) + T_{AR_1} f(X + e_1 + e_2) \end{aligned}$$

Similarly, in this case, the first inequality is true since reject is a feasible action in both states $X + e_1$ and $X + e_2$. Also, the second inequality follows from the supermodularity of f. For Case 3, we have:

Case 3: accept in X, reject in $X + e_1 + e_2$:

$$\begin{aligned} T_{AR_1} f(X + e_1) + T_{AR_1} f(X + e_2) &\leq f(X + e_1) + \frac{c_1}{\beta} + f(X + e_1 + e_2) \\ &= T_{AR_1} f(X) + T_{AR_1} f(X + e_1 + e_2) \end{aligned}$$

Case 4: reject in X, accept in $X + e_1 + e_2$:

$$\begin{aligned}
T_{AR_1}f(X+e_1) + T_{AR_1}f(X+e_2) &\leq f(X+2e_1) + f(X+e_2) + \frac{c_1}{\beta} \\
&\leq f(X) + \frac{c_1}{\beta} + f(X+2e_1+e_2) \\
&= T_{AR_1}f(X) + T_{AR_1}f(X+e_1+e_2).
\end{aligned}$$

In this case, the first inequality is true since accept is a feasible action in state $X + e_1$, and also reject is a feasible action in state $X + e_2$. The second inequality follows by using supermodularity of f twice.

Supermodularity of T_E:

We have two possible cases, which are as follows:

Case 1: accept in $X + e_1 + e_2$:

$$\begin{aligned}
T_E f(X+e_1) + T_E f(X+e_2) &= p_1 f(X+2e_1) + p_2 f(X+e_1+e_2) \\
&+ p_1 f(X+e_1+e_2) + p_2 f(X+2e_2) \\
&= p_1(f(X+2e_1) + f(X+e_1+e_2)) \\
&+ p_2(f(X+e_1+e_2) + f(X+2e_2)) \\
&\leq p_1(f(X+e_1) + f(X+2e_1+e_2)) \\
&+ p_2(f(X+e_1+2e_2) + f(X+e_2)) \\
&= p_1 f(X+e_1) + p_2 f(X+e_2) \\
&+ p_1 f(X+2e_1+e_2) + p_2 f(X+e_1+2e_2) \\
&= T_E f(X) + T_E f(X+e_1+e_2)
\end{aligned}$$

The first inequality is true since we accept in both states $X + e_1$ and $X + e_2$. The second inequality follows from the supermodularity of f, and since the Case 1 accepts in X, and $X + e_1 + e_2$, so $p_1 f(X + e_1) + p_2 f(X + e_2) = T_E f(X)$ and $p_1 f(X + 2e_1 + e_2) + p_2 f(X + e_1 + 2e_2) = T_E f(X + e_1 + e_2)$.

Case 2: reject in $X + e_1 + e_2$:

According to Theorem 4.2, assume $M \geq f(X+2e_1) - f(X+e_1)$ and $M \geq f(X+2e_2) - f(X+e_2)$. We have:

$$p_1 f(X+e_1)+p_2 f(X+e_2)-p_1 f(X+2e_1)-p_2 f(X+e_1+e_2)-p_1 f(X+e_1+e_2)-p_2 f(X+2e_2)$$

$$+M + f(X+e_1+e_2) \geq 0.$$

After rewriting we have:

$$p_1(f(X+e_1) - f(X+2e_1) + M) + p_2(f(X+e_2) - f(X+2e_2) + M) \geq 0,$$

which holds true. Therefore, T_E is supermodular.

Supermodularity of T_{D_i}:

For $i = 1$, operator T_{D1} was defined as:

$$T_{D_1} f(X) = \begin{cases} \frac{\gamma B_1 + x_1}{1+\gamma} f(X-e_1) + \left(\frac{(1+\gamma)B_2 + B_1 - x_1}{1+\gamma}\right) f(X), & if \quad x_1 > B_1, \\ x_1 f(X-e_1) + (B_1 + B_2 - x_1) f(X), & if \quad x_1 \leq B_1. \end{cases}$$

In general, we need to show:

$$T_{D_1} f(X+e_1+e_2) + T_{D_1} f(X) - T_{D_1} f(X+e_1) - T_{D_1} f(X+e_2) \geq 0. \qquad (4.1)$$

In order to check the property, we should examine it for $\{x_1 > B_1, X \in \Psi\}$, $\{x_1 \leq B_1, X \in \Psi\}$ and a boundary like $\{x_1 = B_1, x_2 < B_2\}$.

Case 1 $\qquad x_1 > B_1, x_1 + x_2 \leq B_1 + B_2$

After rewriting property 4.1, we have:

$$\frac{\gamma B_1 + (x_1+1)}{1+\gamma} f(X+e_2) + \left(\frac{(1+\gamma)B_2 + B_1 - (x_1+1)}{1+\gamma}\right) f(X+e_1+e_2)$$

$$+\frac{\gamma B_1 + x_1}{1+\gamma} f(X - e_1) + \left(\frac{(1+\gamma)B_2 + B_1 - x_1}{1+\gamma}\right) f(X)$$

$$-\frac{\gamma B_1 + (x_1+1)}{1+\gamma} f(X) - \left(\frac{(1+\gamma)B_2 + B_1 - (x_1+1)}{1+\gamma}\right) f(X + e_1)$$

$$-\frac{\gamma B_1 + x_1}{1+\gamma} f(X - e_1 + e_2) - \left(\frac{(1+\gamma)B_2 + B_1 - x_1}{1+\gamma}\right) f(X + e_2) \geq 0.$$

By taking the terms in the same parenthesis, we have:

$$\frac{\gamma B_1 + x_1}{1+\gamma} \left(f(X + e_2) + f(X - e_1) - f(X) - f(X - e_1 + e_2)\right)$$

$$+\frac{(1+\gamma)B_2 + B_1 - (x_1+1)}{1+\gamma} \left(f(X + e_1 + e_2) + f(X) - f(X + e_1) - f(X + e_2)\right) \geq 0.$$

Coefficients $\frac{\gamma B_1 + x_1}{1+\gamma}$ and $\frac{(1+\gamma)B_2 + B_1 - (x_1+1)}{1+\gamma}$ are positive, and the remaining terms directly follow the supermodularity of f.

<u>Case 2</u> $x_1 \leq B_1, x_1 + x_2 \leq B_1 + B_2$

According to property 4.1, we have:

$$(x_1+1) f(X + e_2) + (B_1 + B_2 - (x_1+1)) f(X + e_1 + e_2)$$

$$+x_1 f(X - e_1) + B_1 + (B_1 + B_2 - x_1) f(X)$$

$$-(x_1+1) f(X) - (B_1 + B_2 - (x_1+1)) f(X + e_1)$$

$$-x_1 f(X - e_1 + e_2) - (B_1 + B_2 - x_1) f(X + e_2) \geq 0.$$

By taking the terms in the same parenthesis, we have:

$$x_1 \left(f(X + e_2) + f(X - e_1) - f(X) - f(X - e_1 + e_2)\right)$$

$$+(B_1 + B_2 - (x_1+1)) \left(f(X + e_1 + e_2) + f(X) - f(X + e_1) - f(X + e_2)\right) \geq 0.$$

Like the previous case, in this one also coefficients x_1 and $B_1 + B_2 - (x_1 + 1)$ are positive, and the remaining terms directly follow the supermodularity of f.

<u>Case 3</u> $x_1 = B_1, x_2 < B_2 + 1$

After rewriting property (4.1), we have

$$\frac{\gamma B_1 + (x_1+1)}{1+\gamma} f(X+e_2) + \left(\frac{(1+\gamma)B_2 + B_1 - (x_1+1)}{1+\gamma}\right) f(X+e_1+e_2)$$
$$+x_1 f(X-e_1) + (B_1+B_2-x_1) f(X)$$
$$-\frac{\gamma B_1 + (x_1+1)}{1+\gamma} f(X) - \left(\frac{(1+\gamma)B_2 + B_1 - (x_1+1)}{1+\gamma}\right) f(X+e_1)$$
$$-x_1 f(X-e_1+e_2) - (B_1+B_2-x_1) f(X+e_2) \geq 0.$$

Collecting all terms with coefficients x_1, $\gamma(B_1 - x_1)1+\gamma$ and $\frac{(1+\gamma)B_2+B_1-(x_1+1)}{1+\gamma}$ gives:

$$x_1 \left(f(X+e_2) + f(X-e_1) - f(X) - f(X-e_1+e_2)\right)$$
$$+\cancel{\frac{\gamma(B_1-x_1)}{1+\gamma} f(X+e_2)} - \cancel{\frac{\gamma(B_1-x_1)}{1+\gamma} f(X)}$$
$$+\frac{(1+\gamma)B_2 + B_1 - (x_1+1)}{1+\gamma} \left(f(X+e_1+e_2) + f(X) - f(X+e_1) - f(X+e_2)\right)$$
$$+\cancel{\frac{\gamma(B_1-x_1)}{1+\gamma} f(X)} - \cancel{\frac{\gamma(B_1-x_1)}{1+\gamma} f(X+2)} \geq 0.$$

Some terms cancel out, and finally, it reduces to:

$$x_1 \left(f(X+e_2) + f(X-e_1) - f(X) - f(X-e_1+e_2)\right)$$
$$+\frac{(1+\gamma)B_2 + B_1 - (x_1+1)}{1+\gamma} \left(f(X+e_1+e_2) + f(X) - f(X+e_1) - f(X+e_2)\right) \geq 0,$$

which as discussed in the previous cases, is nonnegative. Therefore operator T_{D_1} is supermodular. Also, supermodularity of operator T_{D_2} can be shown like the same procedure discussed above. □

The following proposition establishes a monotone admission policy as stated in the next theorem.

Proposition 4.4. *A class i patient is rejected (accepted) in state $X = (x_i, x_{i^c})$; he/she will be rejected (accepted) in all $X = (x_i, h)$ with $h \geq x_{i^c}$ ($h < x_{i^c}$).*

Proof. For $i = 1$, since the system rejects the type 1 patient in state $X = (x_1, x_2)$, we have

$$f(X) + \frac{c_1}{\beta} \leq f(X + e_1)$$

, so

$$\frac{c_1}{\beta} \leq f(X + e_1) - f(x).$$

It can be easily obtain from supermodularity of f that

$$f(x + e_1 + e_2) - f(x + e_2) \geq f(x + e_1) - f(x).$$

Hence,

$$\frac{c_1}{\beta} \leq f(x + e_1 + e_2) - f(x + e_2).$$

Therefore it would also reject the coming type 1 patient if the state increases to $X = (x_1, x_2 + 1)$. It can be proved for $i = 2$ similarly. □

4.3 Existence of a preferred class

We accept a type i patient at state X if:

$$f(X + e_i) \leq f(X) + \frac{c_i}{\beta}.$$

Equivalently, it means if $f(X + e_i) - f(X)$ is less than a certain value ($\frac{c_i}{\beta}$), the patient is accepted. As a result, finding a proper upper-bound can help us to ensure that a patient type should be always admitted, regardless of the system congestion level. Note that here we assume there is no transfer cost, which means $t_i = 0$ for $i = 1, 2$. To this end, like Ormeci et al. (2001), we define $D_n(ij)(X)$ as the difference in the total expected n-period discounted cost between system

A and system B if system A starts in state X plus one class-i job and system B starts in X plus a class j job. Therefore, we have

$$D_n(10)(X) = V_{n,\beta}(X + e_1) - V_{n,\beta}(X),$$

$$D_n(20)(X) = V_{n,\beta}(X + e_2) - V_{n,\beta}(X),$$

$$D_n(12)(X) = -D_n(21)(X) = V_{n,\beta}(X + e_1) - V_{n,\beta}(X + e_2).$$

Let d_1 and d_{12} be upper-bounds for $D_n(10)$ and $D_n(12)$ for all $x \in \Psi$, and for all n, respectively. We will consider two pairs of coupled systems, one for $D_n(12)$ and the other for $D_n(10)$.

We face three different cases (1) $x_1 < B_1$, $x_2 < B_2$, (2) $x_1 \geq B_1$, $x_2 < B_2$ and (3) $x_1 < B_1$, $x_2 \geq B_2$, from which the greatest upper-bound will be picked as the upper-bound. For case (1) $x_1 < B_1$, $x_2 < B_2$, Let system A' be in state $X + e_1$ and system B' in $X + e_2$. System B' takes the optimal actions and system A' imitates all the actions of system B' in the next transition and afterwards follows the optimal policy. We couple the additional type-2 patient in system B' with the additional type-1 patient as well as all other service and inter-arrival times. So, if type-1 patient leaves the system, which happens with probability μ_1, type-2 patient also leaves. The departure of type-2 patient alone, which happens with probability $\mu_2 - \mu_1$, takes the systems to two different states, $X + e_1$ and X. If there is any other transition, both systems continue to have their additional jobs. For these coupling pairs, letting $V_{n+1}^{A'}(X + e_1)$ be total expected discounted cost of system A', we have:

$$
\begin{aligned}
D_{n+1}(12)(X) &= V_{n+1}(X+e_1) - V_{n+1}(X+e_2) \le V^A_{n+1}(X+e_1) - V_{n+1}(X+e_2) \\
&\le max\Big\{\beta(1-\mu_2)D_n(12)(X) + \beta(\mu_2-\mu_1)D_n(10)(X) + \beta\mu_1 \times 0, \\
&\beta(1-\mu_2)D_n(12)(X) + \beta(\mu_2 - \frac{\mu_1}{1+\gamma})D_n(10)(X) + \frac{\beta\mu_1}{1+\gamma} \times 0, \\
&\beta(1-\frac{\mu_2}{1+\gamma})D_n(12)(X) + \beta(\frac{\mu_2}{1+\gamma} - \mu_1)D_n(10)(X) + \beta\mu_1 \times 0\Big\} \\
&= max\Big\{\frac{\beta(\mu_2-\mu_1)D_n(10)(X)}{1+\beta(\mu_2-1)}, \frac{\beta(\mu_2(1+\gamma)-\mu_1)D_n(10)(X)}{(1+\gamma)(1+\beta(\mu_2-1))}, \\
&\frac{\beta(\mu_1(1+\gamma)-\mu_2)D_n(10)(X)}{\beta(1+\gamma-\mu_2)-(1+\gamma)}\Big\} \\
&= \frac{\beta(\mu_2(1+\gamma)-\mu_1)D_n(10)(X)}{(1+\gamma)(1+\beta(\mu_2-1))}.
\end{aligned}
\tag{4.2}
$$

The same explanation can be used for other cases. We only note that in Case 2 the LOS of the additional type 1 patient has a mean of $\frac{1}{(1+\gamma)\mu_1}$ and in Case 3, the LOS of the additional type 2 patient has a mean of $\frac{1}{(1+\gamma)\mu_2}$.

For $D_{n+1}(10)(X)$, we have two different cases (1) $x_1 < B_1$, (2) $x_1 \ge B_1$, $x_2 < B_2$, which (2) $x_1 \ge B_1$, $x_2 < B_2$ gives the greatest upper-bound. We couple two systems: Let system A be in state $X + e_1$ and system B in X. System A follows the not necessarily optimal policy and rejects all incoming patients in the next arrival transition, and follows the optimal policy afterwards, whereas system B always takes the optimal actions. Consider an arrival. If system B also rejects either of the two types, both systems remain in their current states. Acceptance of a type-2 patient to system B leads two systems to two different states $X + e_1$ and $X + e_2$. If a type-1 patient is admitted to system B, then the two systems couple with no difference in cost. With probability $(1 - \lambda_1 - \lambda_2 - \mu_1)$ there is no departure for additional type 1 patient and with probability μ_1 additional type 1 patient leaves the system. We have:

$$\begin{aligned}
D_{n+1}(10)(X) &= V_{n+1}(X+e_1) - V_{n+1}(X) \\
&\leq max\Bigg\{\beta\{\lambda_1 max\{\underbrace{D_n(10)(X)}_{\text{A and B reject the type 1 patient.}}, \underbrace{\frac{c_1}{\beta}}_{\text{A rejects and B accepts the type 1 patient.}}\} \\
&+ \lambda_2 max\{\underbrace{D_n(10)(X)}_{\text{A and B reject the type 2 patient.}}, \underbrace{\frac{c_2}{\beta} + D_n(12)(X)}_{\text{A rejects and B accepts the type 2 patient.}}\} \\
&+ (1-\lambda_1-\lambda_2-\mu_1)\, D_n(10)(X) + \mu_1 \times 0\}, \\
&\beta\{\lambda_1 max\{D_n(10)(X), \frac{c_1}{\beta}\} \\
&+ \lambda_2 max\{D_n(10)(X), \frac{c_2}{\beta} + D_n(12)(X)\} \\
&+ \left(1-\lambda_1-\lambda_2-\frac{\mu_1}{1+\gamma}\right) D_n(10)(X) + \frac{\mu_1}{1+\gamma} \times 0\}\Bigg\} \\
&= \beta\{\lambda_1 max\{D_n(10)(X), \frac{c_1}{\beta}\} \\
&+ \lambda_2 max\{D_n(10)(X), \frac{c_2}{\beta} + D_n(12)(X)\} \\
&+ \left(1-\lambda_1-\lambda_2-\frac{\mu_1}{1+\gamma}\right) D_n(10)(X)\}.
\end{aligned} \tag{4.3}$$

The following inequality follows from the definition of d_1 and d_{12}.

$$\beta\lambda_1 max\{d_1, \frac{c_1}{\beta}\} + \beta\lambda_2 max\{d_1, \frac{c_2}{\beta} + d_{12}\} + \beta\left(1-\lambda_1-\lambda_2-\frac{\mu_1}{1+\gamma}\right) d_1 \leq d_1.$$

In order to find the minimal d_1, satisfying (4.2) and (4.3) we need to distinguish two cases. We focus on $d_1 \leq \frac{c_1}{\beta}$ to find conditions under which type 1 patients are preferred.

<u>Case 1: $d_1 \leq \frac{c_2}{\beta} + d_{12}$</u> In this case, we have:

$$\lambda_1 c_1 + \beta\lambda_2(\frac{c_2}{\beta} + d_{12}) + \beta\left(1-\lambda_1-\lambda_2-\frac{\mu_1}{1+\gamma}\right) d_1 \leq d_1,$$

so that:

$$d_1 \geq \frac{\frac{\lambda_1 c_1}{\beta} + \lambda_2(\frac{c_2}{\beta} + d_{12})}{\frac{1}{\beta} - 1 + \lambda_1 + \lambda_2 + \frac{\mu_1}{1+\gamma}}.$$

It can be shown that the minimal bounds, are found when both d_1 and d_{12} satisfy the above inequality as equality. Since $d_{12} \leq \frac{\beta(\mu_2(1+\gamma)-\mu_1)d_1}{(1+\gamma)(1+\beta(\mu_2-1))}$, we have:

$$d_1 = \frac{\frac{\lambda_1 c_1}{\beta} + \lambda_2(\frac{c_2}{\beta} + \frac{\beta(\mu_2(1+\gamma)-\mu_1)d_1}{(1+\gamma)(1+\beta(\mu_2-1))})}{\frac{1}{\beta} - 1 + \lambda_1 + \lambda_2 + \frac{\mu_1}{1+\gamma}},$$

from which we set:

$$\underline{d_1} = \frac{(1+\gamma)(c_1\lambda_1(1+\beta(\mu_2-1)) + c_2\lambda_2(1+\beta(\mu_2-1)))}{(1+\gamma)(1+\beta(-2+\lambda_{1+}\lambda_2+\mu_1+\mu_2)) + \beta^2(\lambda_1(\mu_2-1)(1+\gamma)+(\mu_1-1-\gamma)(-1+\lambda_2+\mu_2))},$$

We first check whether $d_1 \leq \frac{c_2}{\beta} + d_{12}$:

$$d_1 \leq \frac{c_2}{\beta} + \frac{\beta(\mu_2(1+\gamma)-\mu_1)d_1}{(1+\gamma)(1+\beta(\mu_2-1))} = \frac{(1+\gamma)(c_2(1+\beta(\mu_2-1)))}{\beta(1+\gamma-\beta(1+\gamma-\mu_1))}.$$

This gives the following condition:

$$c_1 \leq \frac{c_2(1+\gamma+\beta(-1+\gamma(-1+\lambda_1)+\lambda_1+\mu_1))(1+\beta(\mu_2-1))}{\beta\lambda_1(1+\gamma-\beta(1+\gamma-\mu_1))}. \tag{4.4}$$

Now we check if

$$d_1 \leq \frac{c_1}{\beta},$$

which is satisfied if

$$c_1 \geq \frac{\beta(1+\gamma)c_2\lambda_2(1+\beta(-1+\mu_2))}{(1+\gamma-\beta(1+\gamma-\mu_1))(1+\beta(-1+\lambda_2+\mu_2))}. \tag{4.5}$$

Not that we obtain these terms using Wolfram Mathematica since it involves a straightforward but tedious algebra.

Case 2: $\frac{c_2}{\beta} + d_{12} < d_1$

This case gives the following inequality:

$$\lambda_1 c_1 + \beta\lambda_2 d_1 + \beta\left(1 - \lambda_1 - \lambda_2 - \frac{\mu_1}{1+\gamma}\right) d_1 \leq d_1,$$

from which we have:

$$\underline{d_1} = \frac{(1+\gamma)(c_1\lambda_1)}{1 - \beta + \gamma(1-\beta) + \beta\lambda_1(1+\gamma) + \beta\mu_1}.$$

Checking $\frac{c_2}{\beta} + d_{12} < d_1$ gives:

$$\frac{c_2}{\beta} + \frac{\beta(\mu_2(1+\gamma) - \mu_1)d_1}{(1+\gamma)(1+\beta(\mu_2 - 1))} < d_1,$$

so that:

$$d_1 > \frac{(1+\gamma)(c_2(1+\beta(\mu_2-1)))}{\beta(1+\gamma-\beta(1+\gamma-\mu_1))},$$

which is satisfied if

$$c_1 > \frac{c_2(1+\gamma+\beta(-1+\gamma(-1+\lambda_1)+\lambda_1+\mu_1))(1+\beta(\mu_2-1))}{\beta\lambda_1(1+\gamma-\beta(1+\gamma-\mu_1))}. \tag{4.6}$$

Finally,

$$d_1 \leq \frac{c_1}{\beta},$$

if

$$c_1 \geq 0. \tag{4.7}$$

Not that the condition (4.6) is complementary to (4.4), as expected since we check two complementary conditions related to d_1 and d_{12}. Under the given conditions all of the upper-bounds on $D_n(10)(x)$ are non-negative. Thus, it immediately leads to:

Theorem 4.5. *If* $c_1 \geq \frac{\beta(1+\gamma)(c_2\lambda_2(1+\beta(-1+\mu_2)))}{(1+\gamma-\beta(1+\gamma-\mu_1))(1+\beta(-1+\lambda_2+\mu_2))}$, *then type 1 patient is a preferred type for all* $n \geq 0$.

Note that every c_1 that satisfies Equation (4.5), also satisfies Equation (4.7). Therefore, the Theorem 4.5 is feasible. According to Theorem 4.5, by increasing c_2, the threshold also increases, which means the importance (rejection cost) of the type 2 patient, can affect the admission process of type 1 patient.

Assume that system A starts in state $X + e_1$ and system B in $X + e_2$. We couple the additional type 1 patient in system A with the additional type 2 patient as well as all other service and inter-arrival times, so that if type 1 patient leaves the system, type 2 patient also leaves. Let system A follow the optimal policy and B imitate all decisions of A. Then the difference in the expected discounted cost of A and B is always positive, so for patient type 2 we have:

$$D_{n+1}(21)(X) = -D_{n+1}(12)(X) \leq 0 \tag{4.8}$$

For $D_{n+1}(20)(X)$ also we have two different cases (1) $x_2 < B_2$, (2) $x_1 < B_1$, $x_2 \geq B_2$, which (2) $x_1 < B_1$ and $x_2 \geq B_2$ gives the greatest upper-bound. We have

$$\begin{aligned}
D_{n+1}(20)(X) &= V_{n+1}(X + e_2) - V_{n+1}(X) \\
&\leq \beta \Bigg\{ \lambda_1 max\{D_n(20)(X), \frac{c_1}{\beta} + D_n(21)(X)\} \\
&+ \lambda_2 max\{D_n(20)(X), \frac{c_2}{\beta}\} \\
&+ \left(1 - \lambda_1 - \lambda_2 - \frac{\mu_2}{1+\gamma}\right) D_n(20)(X) + \frac{\mu_2}{1+\gamma} \times 0 \Bigg\} \qquad (4.9) \\
&= \beta \Bigg\{ \lambda_1 max\{D_n(20)(X), \frac{c_1}{\beta} + D_n(21)(X)\} \\
&+ \lambda_2 max\{D_n(20)(X), \frac{c_2}{\beta}\} \\
&+ \left(1 - \lambda_1 - \lambda_2 - \frac{\mu_2}{1+\gamma}\right) D_n(20)(X) \Bigg\}.
\end{aligned}$$

The following inequality follows from d_2 being upper-bound.

$$\beta\lambda_1 max\{d_2, \frac{c_1}{\beta} + d_{21}\} + \beta\lambda_2 max\{d_2, \frac{c_2}{\beta}\} + \beta\left(1 - \lambda_1 - \lambda_2 - \frac{\mu_2}{1+\gamma}\right) d_2 \leq d_2.$$

In order to find the minimal d_2, satisfying (4.8) and (4.9) we need to consider the following case.

<u>Case 1: $d_2 \leq \frac{c_2}{\beta}$</u>

$$\beta\lambda_1\left(\frac{c_1}{\beta} + d_{21}\right) + \lambda_2 c_2 + \beta\left(1 - \lambda_1 - \lambda_2 - \frac{\mu_2}{1+\gamma}\right) d_2 \leq d_2.$$

$$d_2 \geq \frac{\frac{\lambda_2 c_2}{\beta} + \lambda_1(\frac{c_1}{\beta} + d_{21})}{\frac{1}{\beta} - 1 + \lambda_1 + \lambda_2 + \frac{\mu_2}{1+\gamma}}.$$

Since $d_{12} \leq 0$, we have:

$$d_2 = \frac{\frac{\lambda_2 c_2}{\beta} + \frac{\lambda_1 c_1}{\beta}}{\frac{1}{\beta} - 1 + \lambda_1 + \lambda_2 + \frac{\mu_2}{1+\gamma}},$$

from which we have:

$$\underline{d_2} = \frac{\lambda_2 c_2 + \lambda_1 c_1}{1 - \beta + \beta\lambda_1 + \beta\lambda_2 + \frac{\beta\mu_2}{1+\gamma}}.$$

which is positive. By checking $d_2 \leq \frac{c_2}{\beta}$, we have:

$$\frac{\lambda_2 c_2 + \lambda_1 c_1}{1 - \beta + \beta\lambda_1 + \beta\lambda_2 + \frac{\beta\mu_2}{1+\gamma}} \leq \frac{c_2}{\beta},$$

which is satisfied if

$$c_2 \geq \frac{\beta(1+\gamma)(c_1\lambda_1)}{1 + \gamma + \beta(-1 + \gamma(\lambda_1 - 1) + \mu_2 + \lambda_1)}. \tag{4.10}$$

Finally, it leads to:

Theorem 4.6. *If $c_2 \geq \frac{\beta(1+\gamma)(c_1\lambda_1)}{1+\gamma+\beta(-1+\gamma(\lambda_1-1)+\mu_2+\lambda_1)}$, then type 2 patient is a preferred type for all $n \geq 0$.*

According to Theorem 4.6, by increasing c_1, the threshold also increases, which means the importance (rejection cost) of the type 1 patient, can directly affect the admission process of type 2 patient.

Remark 4.7. *If $t_i > 0$ for $i = 1, 2$, it can be shown that under certain conditions type i patients are always admitted to their own ward.*

Chapter 5

NUMERICAL RESULTS

5.1 Data Analysis

This is a data analysis provided to analyze the admission and length of stay (LOS) processes. Outputs are based on the computerized records of all admissions to a private hospital over the years 2011 and 2012. Minitab 1.6, MS Excel and Wolfram Mathematica 9 were mainly used to do the statistical analysis. As discussed in chapter 3, we focus on 10 departments.

Daily variability in the arrival pattern is problematic in admission control and impacts cost,access, quality and safety in healthcare delivery. The data show a cyclic daily arrival rate(Figure 5.1). The daily arrival rates do not vary much except for Sunday, on which only emergency operations are performed.

5.1.1 Admission

Table 5.1.1 shows the admission rate to the departments over the period under consideration. Although it has been shown that many arrival processes can be approximated by a Poisson process, here we cannot completely draw such a conclusion. Poisson process does not fit the data with sufficient statistical significance. However, we assume that all arrivals are distributed as a Poisson process.

We perform a Chi Square test to check the arrival distribution. The chi-square statistic is a measure of divergence between your data's distribution and an ex-

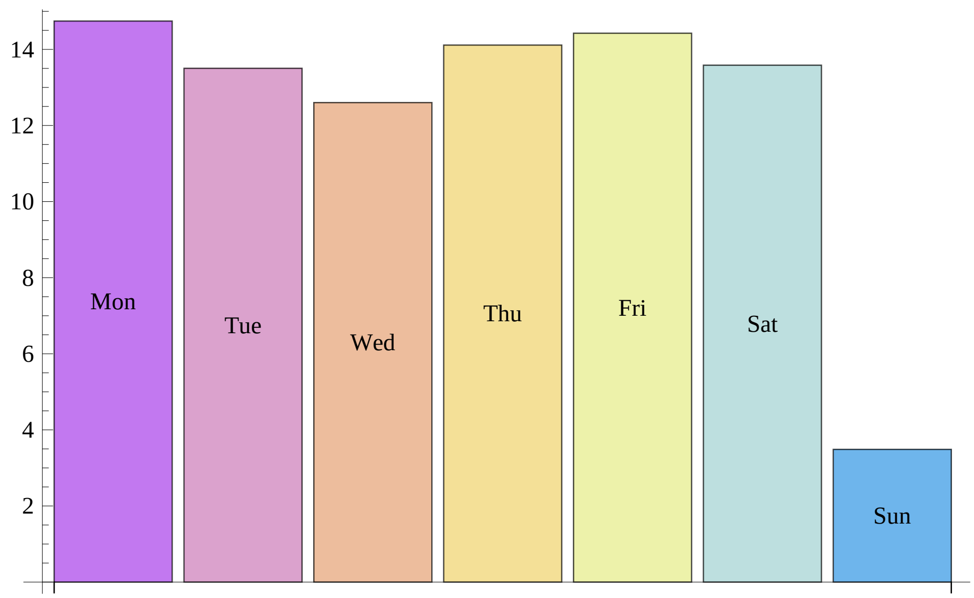

Figure 5.1: Total daily arrival rate to the hospital with respect to the day of the week.

pected or hypothesized distribution of your choice. This test compares the observed values to theoretical values to determine whether there is a significant difference. In this test, p-value is an important factor and helps you determine the significance of your results. P-value greater than the common alpha level of 0.05, suggests that the data follow the Poisson distribution and can be used with analyses that make this assumption.

For example, for Neurosurgery department, high p-value (0.128) supports the claim that the arrival pattern can be approximated by a Poisson process. Figure 5.2 shows that for more than 550 days, there was no arrival and for near 150 days there was just one arrival and for about 30 days there was more than one arrival to the Neurosurgery department. Figure 5.3 shows whether the observations are higher or lower than the expected value. However, for Pediatrics department, a low p-value (0.008) suggests that the data do not follow the Poisson distribution.

Table 5.1: Admission rates to the hospital.

Department	**Arrival rate**	**p-value**
Cardiovascular Surgery	0.683	0.195
ENT	1.006	≈ 0
General Surgery	1.006	≈ 0
Neurosurgery	0.292	0.128
Obstetrics and Gynecology	5.399	0.162
Orthopedics	0.915	0.185
Pediatrics	1.383	0.008
Plastic surgery	0.382	≈ 0
Transplantation	0.203	≈ 0
Urology	1.081	0.039

5.1.2 Length of Stay

In this part we are trying to find a best fit for LOS data. However, finding a comprehensive distribution which precisely explains the data pattern, is a difficult task. If we go through the literature we will see that LOS analysis is one of the more complicated tasks in data analysis. Most researchers used lognormal distribution, whereas some others applied gamma distribution. Although these distributions can be helpful because of their forms, they cannot give a good fit in some conditions. For the sake of simplicity, exponential distribution and Weibull distribution were used to present the LOS in the hospital simulation project.

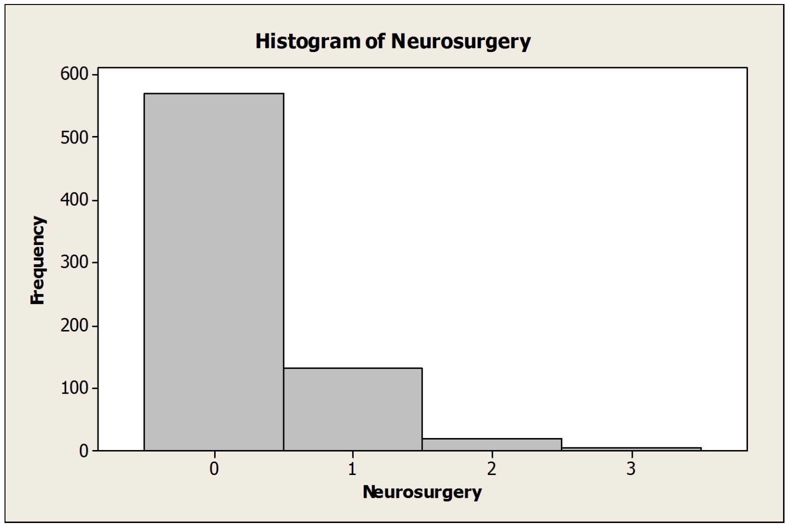

Figure 5.2: Number of arrivals per day for Neurosurgery department

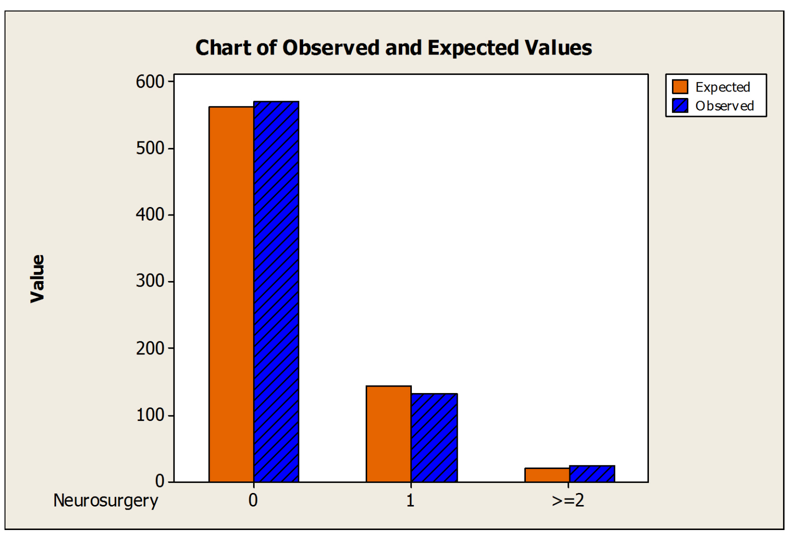

Figure 5.3: Observed and expected values in Chi Square test for Neurosurgery department

However, in order to have a tractable model, we assumed it is exponentially distributed for all of the departments. Table 5.1.2 shows the LOS rates.

Table 5.2: LOS rates.

Department	**LOS rate**
Cardiovascular Surgery	0.364
ENT	1.030
General Surgery	0.376
Neurosurgery	0.243
Obstetrics and Gynecology	0.603
Orthopedics	0.392
Pediatrics	0.730
Plastic surgery	1.049
Transplantation	0.139
Urology	0.689

5.2 Numerical Study

In this section we implement the model on a private hospital. In this hospital, Obstetrics and Gynecology department serves near 40% of arrivals, and is concerned with efficient admission management. Therefore, we group hospital departments into two wards. The second ward contains patients from Obstetrics and Gynecology department, which has 9 beds and the first ward includes the remaining departments (9 departments) with a total of 47 beds. We could not fit the total arrival rate of ward 1's departments with Poisson process; however, we assume the total arrival rate for ward 2 is Poisson. Note that all of these departments can accept patient from each other. The model runs on the following set of parameters:

$\mathfrak{S} = \{\lambda_1 = 6.952,\ \lambda_2 = 5.399,\ \mu_1 = 0.493,\ \mu_2 = 0.603,\ c_1 = 40,\ c_2 = 30,$ $t_1 = 2,\ t_2 = 2.8,\ \beta = 0.9,\ \gamma = 0.05,\ B_1 = 47,\ B_2 = 9,\ M = 0\}$

We use the data to set the arrival and LOS rates. By te way, we used weighted average to find the average LOS for ward 1. For cost parameters c_1 and c_2, we used Helm et al. (2011). We run the model for different cost parameters t_1 and t_2, and finally picked the best ones to use in our model. We analyzed the data to find out how much the LOS rate changes by transferring a patient to the other ward, and set $\gamma = 0.05$. Also, we do not incur a cost when there is no available bed for emergency patient. Moreover, we perform sensitivity analysis on all cost figures as well as γ. Additionally, we change the overall occupancy rate to see its effects.

We use value iteration algorithm to solve the model. Value iteration formally requires an infinite number of iterations to converge exactly to an optimal value. In practice, the algorithm terminates when the value function changes by only a small amount in a sweep. Figure 5.4 shows how $V(40,7)$ converges to the optimal level, when initially $V_0(40,7) = 0$.

One of the basic and quite useful properties of a cost function is its convexity. We are unable to show the convexity of equation (3.1); however we cannot find a counterexample for it. Figures 5.5 and 5.6 show the convexity of $V(x_1,x_2)$ on x_1 and x_2 respectively, and present how the cost function behaves by changing the number of patients in the wards.

Using parameters of set $\mathfrak{S}$, the optimal admission policy for type 1 and 2 patients is presented in Figures 5.7 and 5.8 respectively. According to Figure

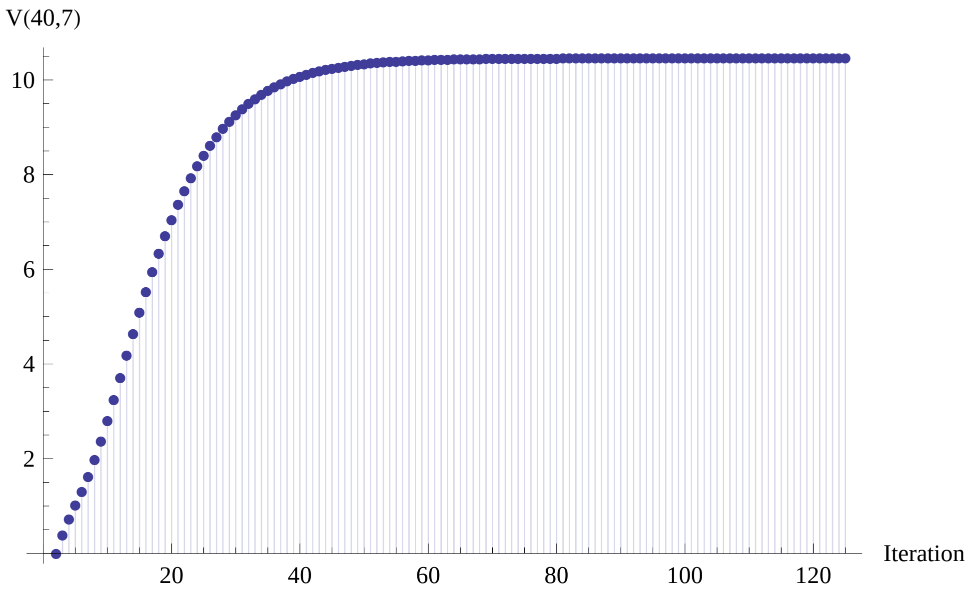

Figure 5.4: Convergence

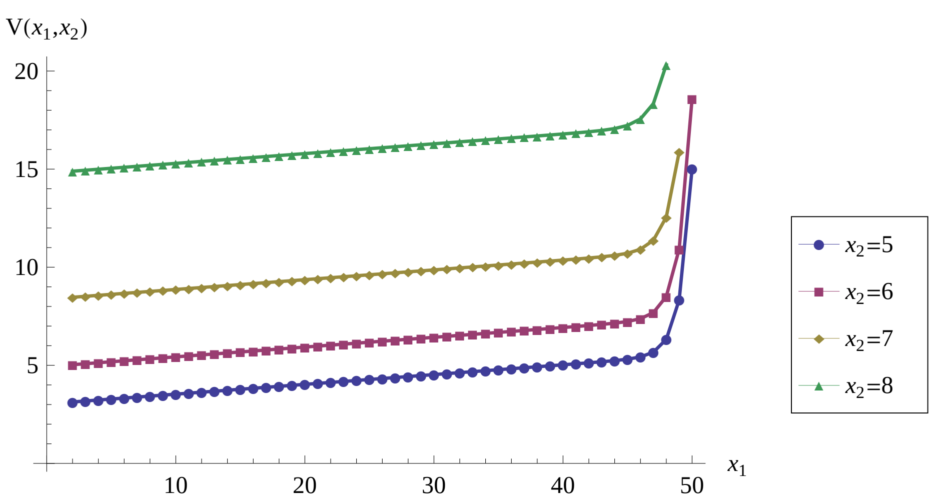

Figure 5.5: Convexity in x_1

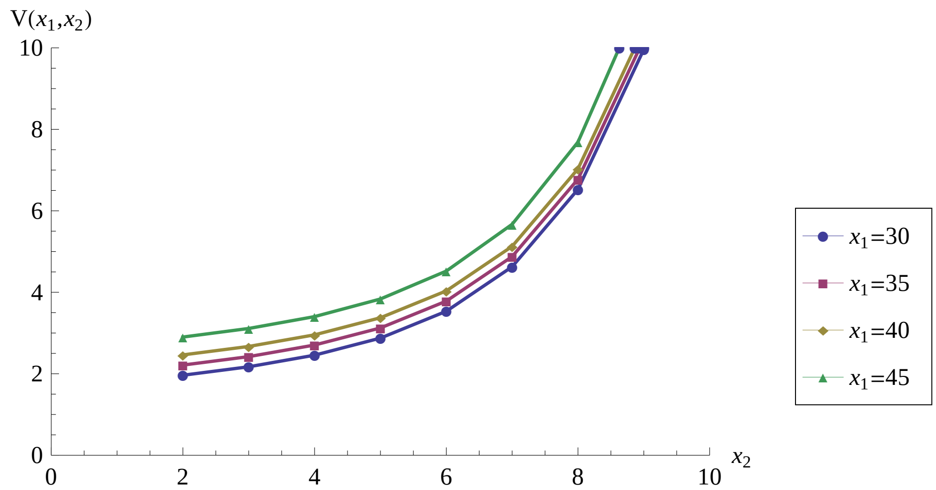

Figure 5.6: Convexity in x_2

5.8, the system rejects the type 2 patient in $X = (0, 53)$, and according to supermodularity of $V(X)$ and Proposition 4.4, it rejects in $X = (k, 53)$ for all $k \geq 1$. However, type 1 patient is always accepted. In the real case, the Obstetrics and Gynecology department always accepts coming patients. Therefore, to have the type 2 patient preferred, we need to increase c_2.

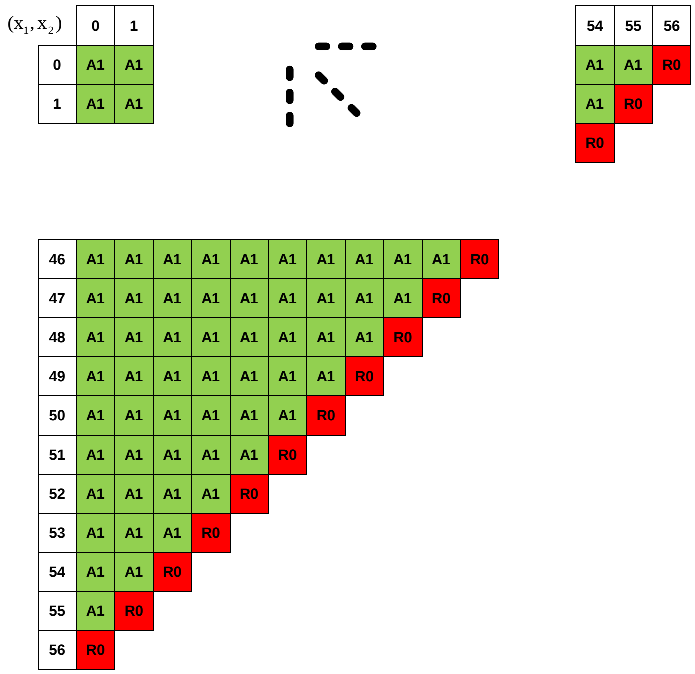

Figure 5.7: Optimal admission policy for patient type 1

5.3 *Sensitivity Analysis*

In this section we want to see how the uncertainty in the output of the model can be apportioned to different sources of uncertainty in its inputs. In heathcare, sensitivity analysis can be useful for a range of purposes, including:

- Testing the robustness of admission policies in the presence of uncertainty,
- Increased understanding of the relationships between input and output variables in the system,
- Searching for errors in the model (by encountering unexpected relationships between inputs and outputs),

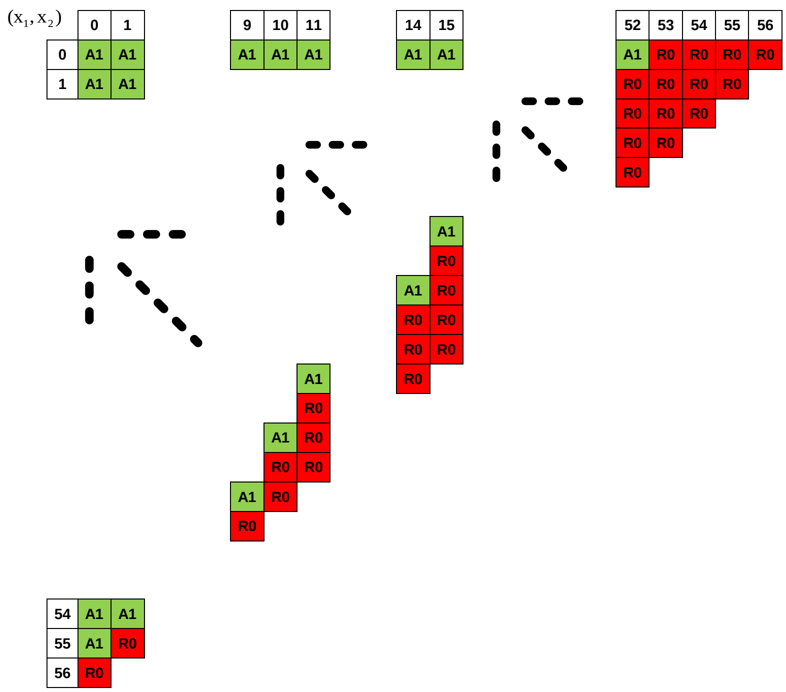

Figure 5.8: Optimal admission policy for patient type 2

- Model simplification by fixing model inputs that have limited effect on the output, or identifying and removing redundant parts of the model structure.

5.3.1 Sensitivity analysis on cost parameters

Cost parameters are definitely the most important ones in the cost-effectiveness analysis of a system. Intuitively, by increasing the cost of rejecting a class-i patient, the system tends to accept the patients of class-i to avoid incurring extra cost. The results presented in Figures 5.9 to 5.12 show that the system accepts class-1 patients in more states when c_1 increases from 10 to 25. Similarly, Figures 5.13 to 5.16 show the same behavior for patient type 2 as c_2 ranges in (10,40). Note that by increasing $c_1(c_2)$, the system rejects type 2(1) patients in more states.

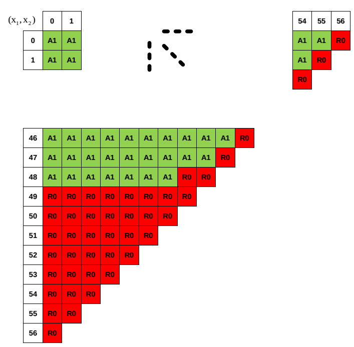

Figure 5.9: Optimal admission policy for type 1 patient ($c_1 = 20$).

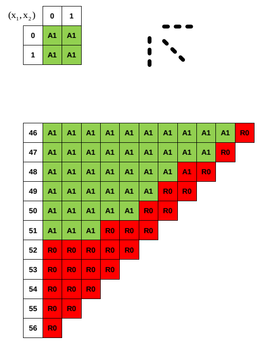

Figure 5.10: Optimal admission policy for type 1 patient ($c_1 = 25$).

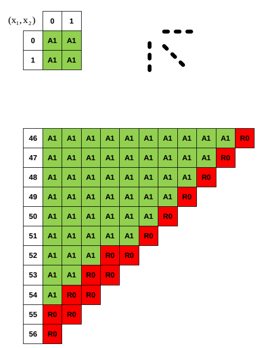

Figure 5.11: Optimal admission policy for type 1 patient ($c_1 = 30$).

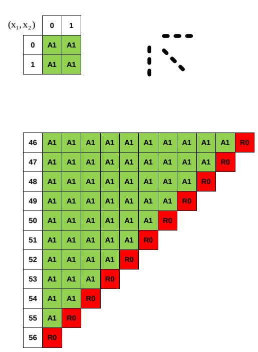

Figure 5.12: Optimal admission policy for type 1 patient ($c_1 = 35$).

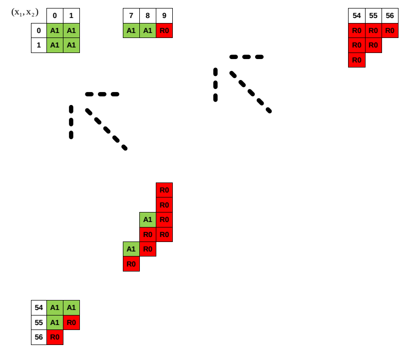

Figure 5.13: Optimal admission policy for type 2 patient ($c_2 = 10$).

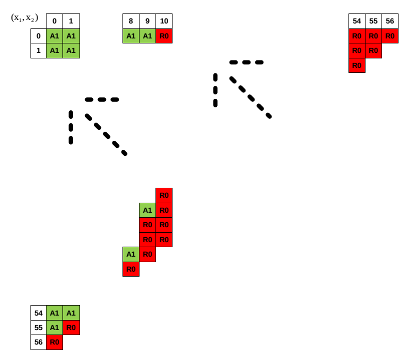

Figure 5.14: Optimal admission policy for type 2 patient ($c_2 = 20$).

The second type of cost in the system is transfer cost. Figures 5.17 to 5.20 show the system rejects patients in more states when the transfer cost t_1 increases from 2 to 11. For low t_1 ($t_1 = 2$), type 1 patients are always accepted. As t_1 increases, the acceptance region shrinks, and it is restricted to the capacity of ward 1 for $t_1 = 11$. Because by increasing transfer cost, the system wants to admit the patient just to his/her specified ward and does not allow the patient to go to the other ward. Figures 5.21 to 5.24 show a similar behavior for patient type 2 as t_2 goes from 1 to 4.

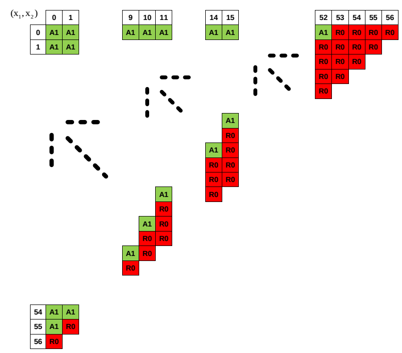

Figure 5.15: Optimal admission policy for type 2 patient ($c_2 = 30$).

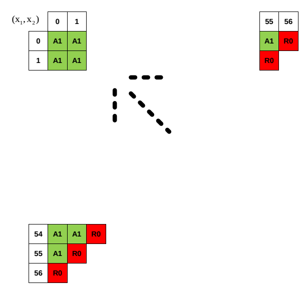

Figure 5.16: Optimal admission policy for type 2 patient ($c_2 = 40$).

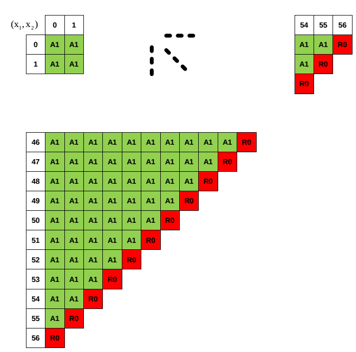

Figure 5.17: Optimal admission policy for type 1 patient ($t_1 = 2$).

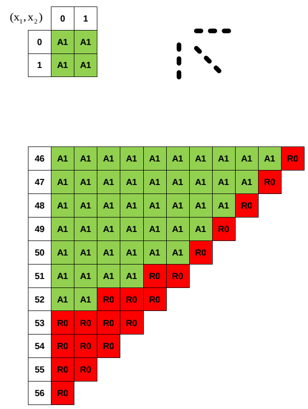

Figure 5.18: Optimal admission policy for type 1 patient ($t_1 = 5$).

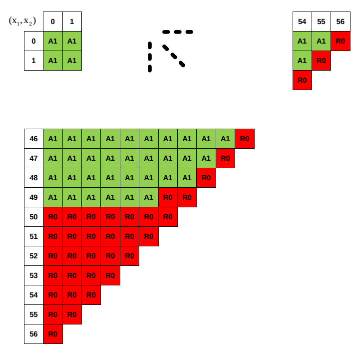

Figure 5.19: Optimal admission policy for type 1 patient ($t_1 = 7$).

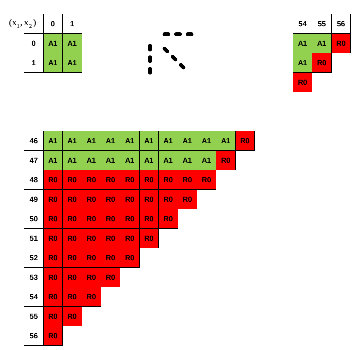

Figure 5.20: Optimal admission policy for type 1 patient ($t_1 = 11$).

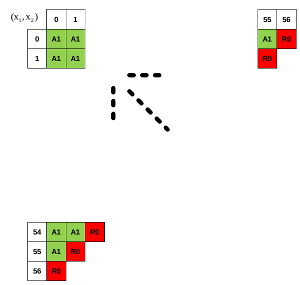

Figure 5.21: Optimal admission policy for type 2 patient ($t_2 = 2$).

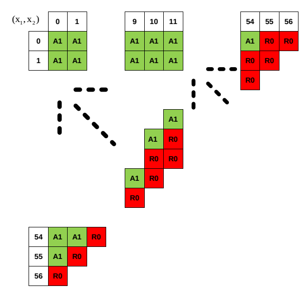

Figure 5.22: Optimal admission policy for type 2 patient ($t_2 = 2.5$).

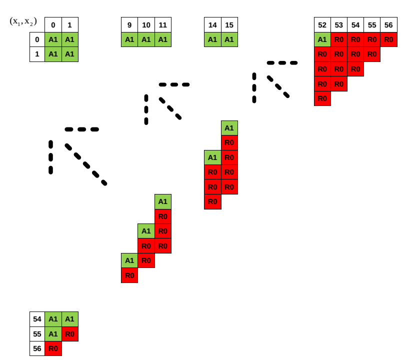

Figure 5.23: Optimal admission policy for type 2 patient ($t_2 = 2.8$).

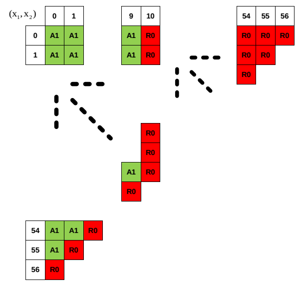

Figure 5.24: Optimal admission policy for type 2 patient ($t_2 = 4$).

5.3.2 Sensitivity analysis on load of the system

In the real case, the load of the system is:

$$O = \frac{\lambda_1 + p_1\lambda_e}{B_1\mu_1} + \frac{\lambda_2 + p_2\lambda_e}{B_2\mu_2} \approx 55\%$$

which is low. In this part we want to increase the patient load, and see how the results change. We fix μ_1, μ_2 and ratio $\frac{\lambda_1}{\lambda_2}$ and change the total arrival rate to increase the occupancy. To observe how the optimal policy for type 1 patient evolves with respect to the load of the system, we run the model for parameter set $\mathfrak{S}$ except for c_1, which is set to 30. Figures 5.25 to 5.28 show that the optimal policy is not very sensitive to the total load of the system. The system rejects slightly more by increasing arrival rates for type 1 patient. Similarly, for type 2 patient, we run the model for parameter set $\mathfrak{S}$. Figures 5.29 to 5.32 show the same behavior for patient type 2.

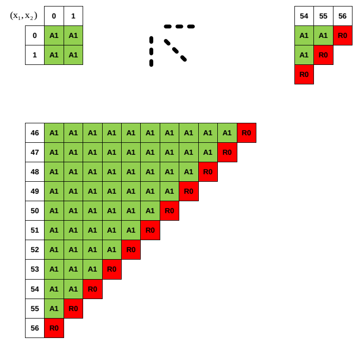

Figure 5.25: Optimal admission policy for type 1 patient ($\lambda_1 = 6.952$, $\lambda_2 = 5.399$ and $O = 55\%$).

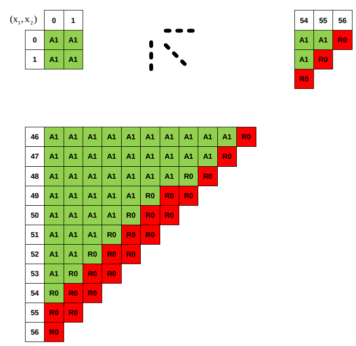

Figure 5.26: Optimal admission policy for type 1 patient ($\lambda_1 = 25.75$, $\lambda_2 = 20$ and $O = 172\%$).

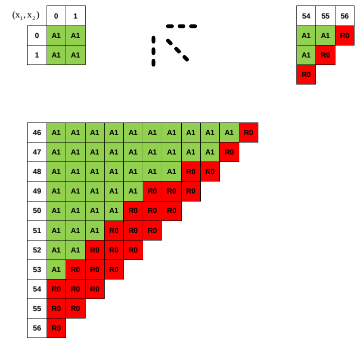

Figure 5.27: Optimal admission policy for type 1 patient ($\lambda_1 = 64.38$, $\lambda_2 = 50$ and $O = 414\%$).

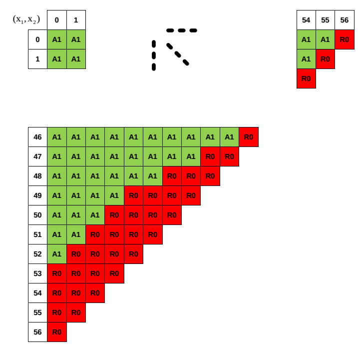

Figure 5.28: Optimal admission policy for type 1 patient ($\lambda_1 = 128.76$, $\lambda_2 = 100$ and $O = 816\%$).

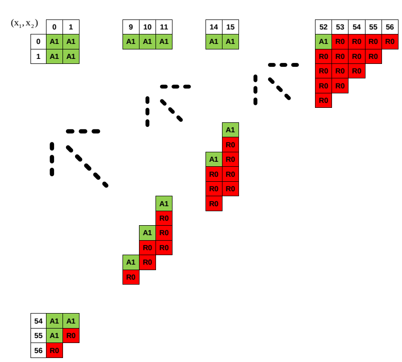

Figure 5.29: Optimal admission policy for type 2 patient ($\lambda_1 = 6.952$, $\lambda_2 = 5.334$ and $O = 55\%$).

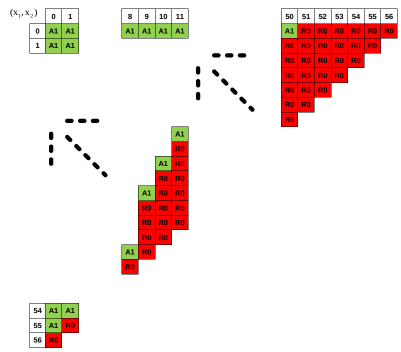

Figure 5.30: Optimal admission policy for type 2 patient ($\lambda_1 = 25.75$, $\lambda_2 = 20$ and $O = 172\%$).

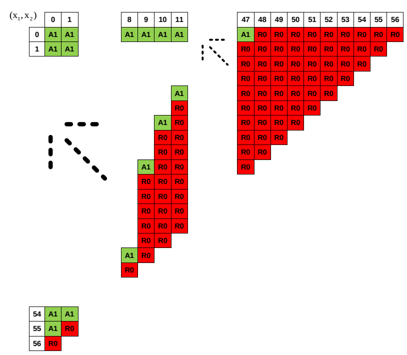

Figure 5.31: Optimal admission policy for type 2 patient ($\lambda_1 = 64.382$, $\lambda_2 = 50$ and $O = 414\%$).

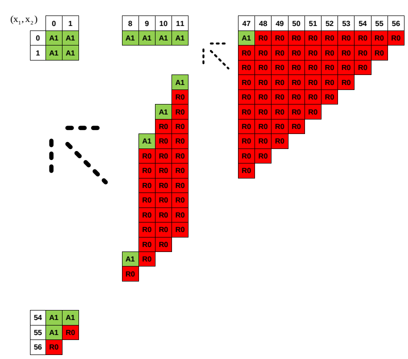

Figure 5.32: Optimal admission policy for type 2 patient ($\lambda_1 = 128.765$, $\lambda_2 = 100$ and $O = 816\%$).

5.3.3 Sensitivity analysis on transfer parameter

The transfers between wards constitute an essential part of this model. Moving patients from one ward to another is vital to ensure that they receive the right care, in the right place at the right time. Any adverse changes in their condition during intra-hospital transport have the potential to be life-threatening. As discussed before, we assumed the LOS rate for transfered patient decreases (he/she stays more). Parameter γ controls this change in the model and here we want to see how the model varies by changing γ. Figures 5.33 to 5.36 show that the system rejects in more states by increasing γ from 0 to 0.2. Because by increasing γ, system does not allow the patient who needs to stay more. The results are based on parameter set $\mathfrak{S}$ except for $c_1 = 30$, which is set to 1. Similarly, Figures 5.37 to 5.40 show the same behavior for patient type 2 as γ goes from 0 to 0.20.

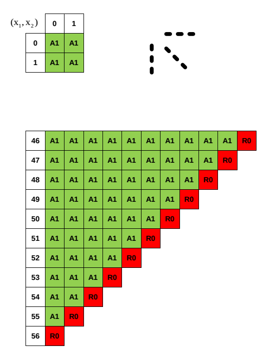

Figure 5.33: Optimal admission policy for type 1 patient ($\gamma = 0$).

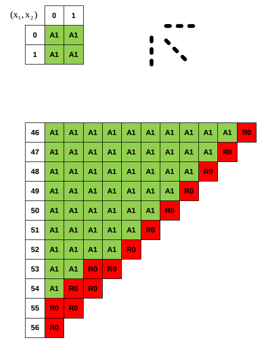

Figure 5.34: Optimal admission policy for type 1 patient ($\gamma = 5\%$).

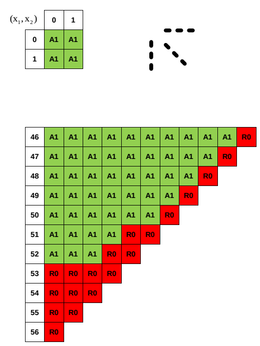

Figure 5.35: Optimal admission policy for type 1 patient ($\gamma = 10\%$).

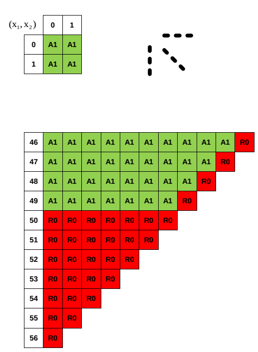

Figure 5.36: Optimal admission policy for type 1 patient ($\gamma = 20\%$).

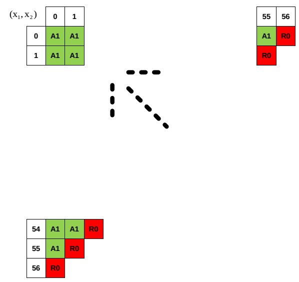

Figure 5.37: Optimal admission policy for type 2 patient ($\gamma = 0$).

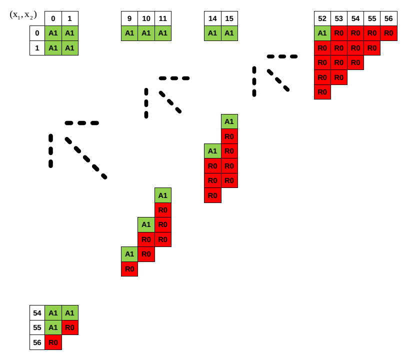

Figure 5.38: Optimal admission policy for type 2 patient ($\gamma = 5\%$).

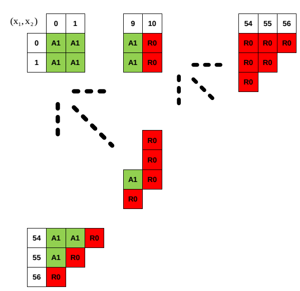

Figure 5.39: Optimal admission policy for type 2 patient ($\gamma = 10\%$).

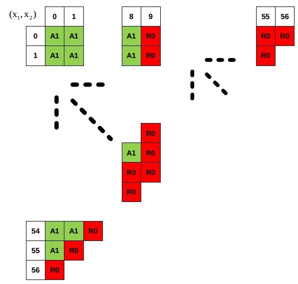

Figure 5.40: Optimal admission policy for type 2 patient ($\gamma = 20\%$).

5.4 A discussion on preferred class

In this section we want to evaluate Theorems 4.5 and 4.6. Again we run the model for parameter set $\mathfrak{S}$ except for $t_1 = 0$, $t_2 = 0$ and $M = 0$. We have:

$$\frac{\beta(1+\gamma)(c_2\lambda_2(1+\beta(-1+\mu_2)))}{(1+\gamma-\beta(1+\gamma-\mu_1))(1+\beta(-1+\lambda_2+\mu_2))} = 32.5863 < c_1 = 40,$$

and

$$\frac{\beta(1+\gamma)(c_1\lambda_1)}{1+\gamma+\beta(-1+\gamma(\lambda_1-1)+\mu_2+\lambda_1)} = 36.4103 > c_2 = 30.$$

Condition 1 is satisfied, so that our results indicate class 1 as preferred. However, Condition 2 is not satisfied, although class-2 patients are preferred. Hence, we cannot always predict the preferred class correctly from these sufficient conditions. Figures 5.41 and 5.42 show the results.

5.5 A discussion on emergency arrival

In previous sections we rejected the emergency patients when there was no free bed. Here, like the real case, we want to always accept the emergency arrival. As a result, we should always have at least one available bed in the hospital. Mathematically, we will incur a large cost if the system became full. We run the model for parameter set $\mathfrak{S}$ except for $M = 1000000$. Figures 5.43 and 5.44

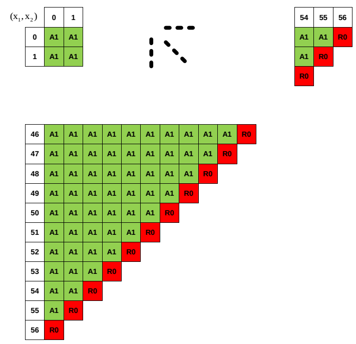

Figure 5.41: Optimal admission policy for type 1 patient (Preferred).

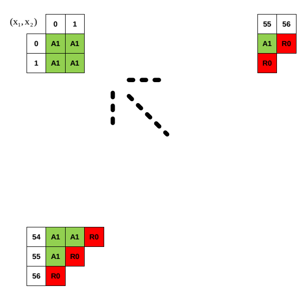

Figure 5.42: Optimal admission policy for type 2 patient (Preferred).

show the results. As we see, we do not have a preferred class in this case since we should always have at least one available bed for emergency patients.

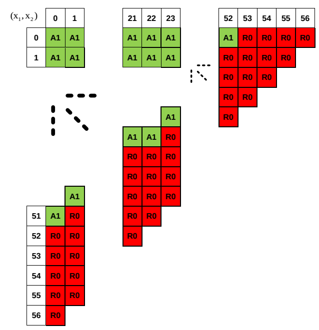

Figure 5.43: Optimal admission policy for type 1 patient.

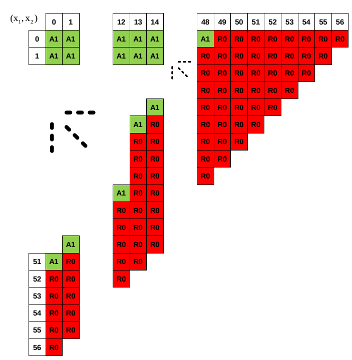

Figure 5.44: Optimal admission policy for type 2 patient.

Chapter 6

CONCLUSION

Bed management is a core operational activity in all hospitals; however, many hospitals have problems with it. In healthcare, bed management systems are designed to help hospitals control patient flow and meet performance targets. In many hospitals and health centers, admission departments are often overcrowded with patients waiting for available rooms. Poor bed management causes some major impacts on patients like waiting too long for treatment, being diverted to other hospitals or transferred to other wards and cancellations.

This research develops an analytical tool for managing admissions to reach a smooth workload, based upon the elective patient flow at a private hospital in Istanbul, Turkey. We have taken the penalty costs as would be lost profit if the patient is turned away or transferred from the department. Here we allow two hospital wards to interact with each other and develop a set of sufficient conditions which ensure that a patient type should be admitted or not, regardless of the system congestion level. The MDP model develops a new method for allocating patients to hospital wards, which helps the hospital

- To decide whether to accept a patient or not,
- To allocate patients to their specified wards,
- To decrease bed occupancy and make the admission process smoother,

- To increase productivity and generate higher revenue.

This model has been evaluated through a real dataset. The model looks at the hospital on a strategic rather than a tactical level. It does not focus on the detailed mechanisms that cause the overall patterns of admissions. The model attempts to seek patterns directly from the data and hospital work-flow to convert it into a mathematical structure. The basic simplicity of the model makes it very flexible and also portable. It could be applied to different wards of a hospital as well. According to admission parameters, hospital managers can find threshold policies, and consequently they can decide whether to accept a patient or not. In long run, this will affect the performance of hospital and can reduce the occupancy rate in emergency departments and admission gateways.

For future work, we should apply the model to other hospitals and different settings to test the variance, and consistency of the parameters. We should also compare the results with other models. Adding more patient classes and waiting lists can also make the model more feasible.

BIBLIOGRAPHY

Altman, E., Jimenez, T., and Koole, G. (1998). On optimal call admission control. *In Proceedings of the IEEE Conference on Decision and Control*.

Arenas, M., Bilbao, A., Caballero, R., Gomez, T., Rodriguezand, M., and Ruiz, F. (2002). Analysis via goal programming of the minimum achievable stay in surgical waiting lists. *Journal of the Operational Research Society*, 53:387 – 396.

Bekker, R. and Koeleman, P. M. (2011). Scheduling admissions and reducing variability in bed demand. *Health Care Manag Science*, 14(3):237 – 249.

Belien, J. and Demeulemeester, E. (2007). Building cyclic master surgery schedules with leveled resulting bed occupancy. *European Journal of Operational Research*, 176(2):1185 – 1204.

Blakea, J. T. and Carterb, M. W. (2002). A goal programming approach to strategic resource allocation in acute care hospitals. *European Journal of Operational Research*, 140(3):541 – 561.

Bruin, A. M. D., Bekker, R., Zanten, L. V., and Koole, G. M. (2010). Dimensioning hospital wards using the erlang loss model. *Annals of Operations Research*, 178(1):23–43.

Carrizosa, E., Conde, E., and Munoz-Marquez, M. (1998). Admission policies in loss queueing models with heterogeneous arrivals. *Management Science*, 44(3):311 – 320.

de Bruin, A. M., van Rossum, A. C., Visser, M. C., and Koole, G. M. (2007). Modeling patient flows using a queuing network with blocking. *Health Care Management Science*, 10(2):125 – 137.

Dobson, G., Hasija, S., and Pinker, E. J. (2011). Reserving capacity for urgent patients in primary care. *European Journal of Operations Research*, 20(3):456 – 473.

Fatovich, D. M., Nagree, Y., and Sprivulis, P. (2005). Access block causes emergency department overcrowding and ambulance diversion in perth, western australia. *Emergency Medicine Journal*, 22(5):351 – 354.

Forster, A. J., Stiell, I., Wells, G., Lee, A. J., and van Walraven, C. (2003). The effect of hospital occupancy on emergency department length of stay and patient disposition. *Academic Emergency Medicine*, 10(2):127 – 133.

Gerchak, Y., Gupta, D., and Henig, M. (1996). Reservation planning for elective surgery under uncertain demand for emergency surgery. *Management Science*, 42:321 – 334.

Ghoneim, H. and S. Stidham, J. (1985). Control of arrivals to two queues in series. *European Journal of Operational Research*, 21(3):399 – 409.

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

Gorunescu, F., McClean, S. I., and Millard, P. H. (2002b). Using a queuing model to help plan bed allocation in a department of geriatric medicine. *Health Care Management Science*, 5(4):307 – 312.

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

Hajek, B. (1984). Optimal control of two interacting service stations. *IEEE Transactions on Automatic Control*, 29(6):491 – 499.

Han, J. H., Zhou, C., France, D. J., Zhong, S., Jones, I., Storrow, A. B., and Aronsky, D. (2007). The effect of emergency department expansion on emergency department overcrowding. *Academic Emergency Medicine*, 14(4):338 – 343.

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

Harper, P. R. and Shahani, A. K. (2002). Modelling for the planning and management of bed capacities in hospitals. *Journal of the Operational Research*, 53(1):11 – 18.

Harrison, G. W., , Shafer, A., and Mackay, M. (2005). Modelling variability in hospital bed occupancy. *Health Care Management Science*, 8(4):325 – 334.

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

Hoot, N. R. and Aronsky, D. (2008). Systematic review of emergency department crowding: Causes, effects, and solutions. *Annals of Emergency Medicine*, 52(2):126 – 136.

Hordijk, A. and Koole, G. (1992). On the assignment of customers to parallel queues. *Probability in the Engineering and Informational Sciences*, 6(4):495 –511.

Jun, J. B., Jacobson, S. H., and Swisher, J. R. (1999). Application of discrete-event simulation in health care clinics: A survey. *Journal of the Operational Research*, 50(2):109 – 123.

Kao, E. P. C. and Tung, G. G. (1981). A queuing model for bed-occupancy management and planning of hospitals. *Management Science*, 27(5):507 – 520.

Koole, G. (1998). Structural results for the control of queueing systems using event-based dynamic programming. *Queueing Systems*, 30(3):323–339.

Ku, C. Y. and Jordan, S. (1997). Access control to two multiserver loss queues in series. *IEEE Transactions on Automatic Control*, 42(7):1017 – 1023.

Lewis, M., Ayhan, H., and Foley, R. (1999). Bias optimality in a queue with admission control. *Probability in the Engineering and Informational Sciences*, 13(3):309 – 327.

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

Lowery, J. C. (1996). Design of hospital admissions scheduling system using simulation. *WSC 96 Proceedings of the 28th conference on Winter simulation*, pages 1199 – 1204.

Machlin, S. R. and Carper, K. (2007). Expenses for hospital inpatient stays, 2004. *Statistical Brief 164*.

McManus, M. L., Long, M. C., Cooper, A., and Litvak, E. (2004). Queuing theory accurately models the need for critical care resources. *Anesthesiology*, 100(5):1271 – 1276.

Miller, B. L. (1969). A queueing reward system with several customer classes. *Management Science*, 16(3):234 – 245.

N. Koizumi, E. Kuno, T. E. S. (2005). Modeling patient flows using a queuing network with blocking. *Health Care Management Science*, 8(1):49 – 60.

NHS (1992). *Lying in Wait: the Use of Medical Beds in Acute Hospitals.* Audit Commission/HMSO.

Ormeci, L. (2004). Dynamic admission control in a call center with one shared and two dedicated service facilities. *IEEE Transactions on Automatic Control*, 49(7):1157–1161.

Ormeci, L., Burnetas, A., and van der Wal, J. (2001). Admission policies to a two-class loss system. *Stochastic Models*, 17(4):513 – 540.

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

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

Puterman, M. (1994). *Markov Decision Processes.* John Wiley and Sons, New York.

Roland, B., di Martinelly, C., and Riane, F. (2006). Operating theatre optimization: A resource-constrained based solving approach. *Service Systems and Service Management, International Conference*, 1:443 – 448.

Ross, K. W. (1995). *Multiservice Loss Models for Broadband Telecommunication Network*. Springer-Verlag, Great Britain.

Savin, S. V., Cohen, M. A., Gans, N., and Katalan, Z. (2005). Capacity management in rental businesses with two customer bases. *Operations Research*, 53(4):617–631.

Schaefer, A. J., Bailey, M. D., Shechter, S. M., and Roberts, M. S. (2004). Modeling medical treatment using markov decision processes. *Journal of the American Hospital Association*, 70:593–612.

Sprivulis, P. C., Silva, J. D., Jacobs, I. G., Frazer, A. R. L., and Jelinek, G. A. (2006). The association between hospital overcrowding and mortality among patients admitted via western australian emergency departments. *Medical Journal of Australia*, 184(5):208 – 212.

Trzeciak, S. and Rivers, E. P. (2003). Emergency department overcrowding in the united states: an emerging threat to patient safety and public health. *Emergency Medicine Journal*, 20(5):402 – 405.

VanBerkel, P. T. and Blake, J. T. (2007). A comprehensive simulation for wait time reduction and capacity planning applied in general surgery. *Health Care Management Science*, 10(4):373 – 385.

Vissers, J. M. H., Adan, I. J. B. F., and Bekkers, J. A. (2005). Patient mix optimization in tactical cardiothoracic surgery planning: a case study. *IMA Journal of Management Mathematics*, 16(3):281 – 304.

Young, J. P. (1965). Stabilization of inpatient bed occupancy through control of admissions. *Journal of the American Hospital Association*, 39:4148.