Submitted to *Operations Research*
manuscript OPRE-2014-02-100.R3

Managing Patient Admissions in a Neurology Ward

Saied Samiedaluie
Alberta School of Business, University of Alberta, Edmonton, Alberta T6G 2R6, Canada
samiedal@ualberta.ca

Beste Kucukyazici, Vedat Verter
Desautels Faculty of Management, McGill University, Montreal, Quebec H3A 1G5, Canada
beste.kucukyazici@mcgill.ca, vedat.verter@mcgill.ca

Dan Zhang
Leeds School of Business, University of Colorado Boulder, Boulder, Colorado 80309, USA
dan.zhang@colorado.edu

We study patient admission policies in a neurology ward where there are multiple types of patients with different medical characteristics. Patients receive specialized care inside the neurology ward and delays in admission to the ward will have negative impact on their health status. The level of this impact varies among patient types and depends on the severity of patients. Patients are also different in terms of arrival rate and length of stay at the ward. The patients normally wait in emergency department until a ward bed is assigned to them. We formulate this problem as an infinite-horizon average cost dynamic program and propose an efficient approximation scheme to solve large-scale problem instances. The computational results from applying our model to a neurology ward show that dynamic policies generated by our approach can reduce the overall deterioration in patients' health status compared to several alternative policies.

Key words: Patient Admission, Neurology Ward, Approximate Dynamic Programming
History: This paper was first submitted on February 19, 2014 and was accepted on September 24, 2016 after two revisions.

1. Introduction

Neurological diseases, including Alzheimer, Amyotrophic lateral sclerosis (ALS), Multiple Sclerosis, Spinal Cord Injury and Stroke, represent leading causes of death and disability in the Canadian and US populations (World Health Organization, 2006). Many neurological conditions are chronic, worsen over time and produce a range of functional limitations posing daily challenges to the patients and their caregivers. For example, Heart and Stroke Foundation (`www.heartandstroke.com`) identifies stroke as the third leading cause of death in Canada with about 14,000 fatalities each year; and reports that about 300,000 Canadians are living with the

effects of stroke. The Global Burden of Disease study conducted in 2002 by the World Health Organization also determined that neurological conditions accounted for 38.3% of the disability-adjusted life years worldwide (Lopez et al., 2006), while the percentages observed in developed countries are much higher than the global average. The neurological conditions also have an economic burden. The incidence of most neurological diseases increases with age, and this is a particular concern for healthcare providers and policy makers in an era of aging population. The latest estimate by the Canadian Institute for Health Information (`www.cihi.ca`) pertaining to the total cost of neurological illnesses in Canada is $185 billion annually, 55% of which is direct cost. With respect to utilization of hospital-based services, around 10% of acute care hospitalizations and 20% of patient days in acute care hospitals on average include patients with one of the neurological conditions, including the secondary diagnosis.

Very few neurological conditions are fully curable. In the event of an acute episode, the neurology patients are admitted to the hospital through the emergency departments (EDs). Diagnosis of such conditions in ED requires extensive physical examinations, brain imaging (CT or MRI) and other diagnostic tests. Following these diagnostic tests, the hospital admission decision is made by a neurologist. Recent studies have shown that such critically ill patients are more effectively treated in specialized inpatient settings, i.e., neurology wards, offering properly organized care (Chalfin et al., 2007, Collaboration, 2013). The features of a neurology ward include the care given by a specialized nursing team, the use of extensively equipped beds, the availability of occupational, speech and physical therapies as well as social workers (Stroke Unit Trialists' Collaboration, 2007). As a result some patients' quality-adjusted life years can be improved significantly through enhanced functional abilities. The accessibility to such specialized care is particularly time-sensitive for patients with acute conditions (Castillo, 1999). Indeed, Kucukyazici et al. (2010) observed that the potential benefits of specialized care might be offset by long delays in ED prior to admission to a neurology ward. To avoid such situations the neurologist may find it necessary to transfer patients to another hospital. This is a decision neurology ward managers strive to avoid since the patient faces additional waiting time at the transfer destination.

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

In many neurology wards, a static patient admission policy is used by assigning a fixed number of beds to each type of disease. Sometimes, a certain number of beds are used as flexible beds and shared among different types of patients. For example, at the Montreal Neurological Hospital, there are sixteen beds in the neurology ward, where six beds are dedicated to stroke patients, six beds are dedicated to non-stroke neurology patients and four of them are used as flexible beds to admit either stroke or non-stroke patients.

In this paper, we focus on patient admissions from the ED that involve the development of *rules* for the allocation of inpatient beds among multiple types of patients as well as the patient transfers. In designing such admission policies, the physicians face the trade-off between (i) the higher risk of deteriorated functionality due to extended ED stays for more severe patients and (ii) the increased risk of blocking due to longer length of stays of these patients. An additional trade-off is between the benefits of reducing the ED boarding time by transferring the patient to another hospital and the inconvenience associated with the transfer. To address these trade-offs, we formulate an infinite-horizon average cost dynamic program (DP) and propose an efficient approximation scheme to solve large-scale problem instances. Our objective is to minimize the average opportunity cost of waiting and transferring by finding the most appropriate patient admission policy from the ED.

To the best of our knowledge, this is the first paper that makes an explicit effort to model the differentiating features of neurology wards, and hence provides managerial insights specific to this domain. Our contributions are three-fold. First, from a modeling perspective, we recognize the significance of the presence of a specialized team of care providers in neurology wards, which renders off-servicing policies infeasible for neurology patients. In dealing with the hard capacity constraints, we incorporate the possibility of patient transfers to other hospitals that are not well studied in the prevailing literature. Second, from the viewpoint of methodology, we develop an LP-based approximate dynamic programming (ADP) approach. While this method typically involves a large-scale LP (e.g., de Farias and Van Roy (2006)), our approach involves solving a number of small DPs that are derived by employing a non-linear functional approximation. We tackle the subsequent complexity by a novel decomposition that results in smaller DPs. We also develop an ADP-based Priority Cut-off policy that not only performs well by incorporating the state of the system in making the patient admission decisions, but also is easy to implement. Lastly, on the managerial side, we highlight the weaknesses of the static patient admission and ad-hoc patient transfer policies that are currently popular. In particular, we show that by incorporating the current utilization of the ward and the nature of the waiting line, it is possible to achieve lower costs and better trade-offs between waiting times and patient transfers.

The remainder of the paper is organized as follows: We provide an overview of the most relevant literature in Section 2. The DP formulation is provided in Section 3, whereas we provide an overview

of the data set from Montreal Neurological Hospital as well as the methods used for estimating the model parameters in Section 4. The properties of the optimal policy are discussed in Section 5. In Section 6, the solution methodology is presented. Section 7 provides some numerical examples to compare the policies obtained through ADP approach to other admission policies. We provide some concluding remarks in Section 8.

2. Literature Review

The patient admission problem has received attention in the academic literature for more than four decades. Among the first studies, Kolesar (1970) develops a Markovian model that incorporates the scheduling of outpatients as well as the admission of inpatients that need immediate hospitalization. Esogbue and Singh (1976) considers the admission problem for two types of patients with the following objectives: maximization of occupancy and minimization of unsatisfied requests. They develop a birth and death process based on a priority cut-off policy and solve for the optimal value of cut-off priority.

Lapierre et al. (1999) develops a time-series model based on hourly census data that assists with the allocation of beds between different medical units within a hospital. Using this model, hospital administrators can decide how many beds should be allocated to each unit to have the same number of bed shortage occurrences across the units. Li et al. (2008) presents an integrated model of queueing and goal programming (GP) that is illustrated through allocation of beds across the departments of a hospital in China. A queueing model is used to compute certain performance measures of the system, for example, patient admission probability. The GP methodology is used to construct a multi-objective decision model taking into account the targets and objectives of hospital management and department heads.

The recent paper of Ayvaz and Huh (2010) that studies allocation of hospital capacities among different types of patients shares some features with our work. They consider two types of patients: type-1 patients who arrive at the system and wait until they are served and type-2 patients who leave the system if they are not immediately accommodated (i.e., balking patients). They assume that each patient requests only one unit of capacity and whenever she is admitted, she stays only until the end of that day irrespective of time of admission. This means that at the beginning of each day, all the capacity becomes available. A discounted total cost dynamic program is developed to find the optimal number of admissions per day. To solve the model, they propose a heuristic policy that protects some portion of capacity for type-2 patients. Helm et al. (2011) incorporates the existence of an expedited patient queue which includes those patients that need to be seen within a few days. They optimize the admission threshold policy by formulating the problem as a Markov decision process (MDP).

From a methodological perspective, the papers that use approximate dynamic programming for patient scheduling and admission problems are relevant to our work. Green et al. (2006) considers capacity allocation of a diagnostic medical facility among different types of patients. They develop a finite-horizon DP which is approximated using linear value functions and a heuristic policy is generated based on this linear approximation. Patrick et al. (2008) formulates advance scheduling of patients with multiple priorities for a diagnostic facility as a discounted infinite-horizon MDP. By considering an affine value function approximation, they produce an approximate linear program (ALP), which is solved by applying column generation technique on its dual problem. Using the solution of the ALP, they develop a booking policy and present the optimality gaps. The same approach is used by Sauré et al. (2012) to schedule cancer patients for radiation therapy sessions. These types of patients require more than one appointment over the planning horizon while Patrick et al. (2008) assumes each patient requires only one appointment.

There is a large number of papers in other areas such as production planning and scheduling, revenue management and communication networks that were pertinent to our work. Carr and Duenyas (2000), De Vericourt et al. (2002) and Paschalidis and Tsitsiklis (2000) are good examples of such papers with relevant modeling and methodological components. In the interest of space, our review is confined to the healthcare domain. Before we turn to the model statement, it is important to highlight the differentiating characteristics of our work. First, all types of patients can wait for service as long as there is space available in the waiting area (i.e., ED). Second, we incorporate the decision about transferring the patients to another hospital. Third, we also consider the different LOSs associated with different patient types. The resulting decision is rather complex from the analytical viewpoint. Hence, we combine queueing methods and approximate dynamic programming (ADP) in devising an integrated solution procedure.

3. The Dynamic Programming Formulation

We consider the problem of admitting patients with different clinical conditions into a neurology ward. There are n types of patients indexed by $i \in \{1, \ldots, n\}$ where type 1 is the least severe patient and type n is the most severe patient. There are B beds available in the ward. We assume that the beds are multi-purpose, i.e., each bed can be used for admitting the patient irrespective of her neurological condition. Patients usually wait in the ED before a bed in the ward is assigned to them. It is generally undesirable to keep neurology patients in the ED due to the lack of the special care needed by this group of patients. The health status of a patient with severe condition deteriorates much faster than one with a non-severe condition, in response to delays in admission to the ward. Assuming that dis-utilities associated with such delays can be expressed in quality of

life related terms, let $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_n)^{\mathrm{T}}$ denote the waiting cost vector, where π_i is the waiting cost per unit time for a patient of type i. Note that $\pi_i \leq \pi_j$ for $i < j$.

We assume type-i patients arrive according to a Poisson process with the rate of λ_i patients per unit time. Upon the arrival of a new patient, the ward manager decides whether to accept or transfer the patient to another hospital. Transferring a type-i patient to another hospital incurs a lump–sum cost, denoted by κ_i. Let $\boldsymbol{\kappa} = (\kappa_1, \kappa_2, \ldots, \kappa_n)^{\mathrm{T}}$ be the transfer cost vector. If the patient is accepted, she is either given a bed or joins the queue and waits until a bed becomes available for her. Whenever a type-i patient is admitted to a bed, we assume she occupies the bed for a time which is exponentially distributed with mean of μ_i^{-1} (which is also called average LOS). Consequently, μ_i indicates the discharge rate for patients of type i. For patients with the same disease, the average LOS for more severe patients tends to be longer. We assume that arrivals and discharges occur independently from each other. When a patient is discharged, a decision is made on whether to admit a patient from the queue to the ward. The decision-making process should be based on the number of waiting patients from each type and also the number of empty beds available.

To find the best admission policy, we formulate the problem as a continuous-time dynamic program. This enables us to limit our attention only to those times when there is a change in the state of the system (Puterman, 1994). The change in the state of the system can be either an arrival of a patient or a discharge of a patient from the ward. The time horizon is considered to be infinite which is consistent with the idea of running a hospital ward. This problem can be formulated either as a total discounted cost or an average cost model. While the total discounted approach seems easier to apply, the dependency of the optimal policy on the discount factor and initial state is a major drawback. Thus, we use an average cost dynamic program, where the objective is to minimize the long-run average cost of the system.

3.1 State Variables

The state of the system includes information about the number of waiting patients and the number of occupied beds by each patient type. We need to distinguish between the beds occupied by different patient types because the discharge rates are not the same for different types. Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}}$, where x_i is the number of waiting type-i patients, and $\mathbf{b} = (b_1, b_2, \ldots, b_n)^{\mathrm{T}}$, where b_i is the number of beds occupied by type-i patients. The state of the system is given by $(\mathbf{x}, \mathbf{b})$. Note that $\mathbf{x}$ and $\mathbf{b}$ are n-dimensional column vectors. We assume the total number of waiting neurology patients is constrained by K, which reflects the hospital's policy with regards to the quality of care. Due to time sensitivity for stabilizing neurology patients as soon as possible, hospitals prefer not to have these patients boarding at the ED for extended period of time. Hence, we have $\sum_{i=1}^{n} x_i \leq K$.

At any time, at most B patients are in the beds, i.e., $\sum_{i=1}^{n} b_i \leq B$. So the state space is finite. We use *post-action* state variables so that the transition rate depends only on the state of the system but not on the actions.

3.2 Actions

Since we model the problem as a continuous time dynamic program, the moments that we make a decision are restricted to those times that the state of the system changes (Puterman, 1994). We classify the possible actions based on the cause of state changes.

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

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

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

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

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

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

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

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

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

$$\mathcal{D}_i(\mathbf{x}) = \left\{ (d_{i1}, \ldots, d_{in}) \in \{0,1\}^n \middle| d_{ij} \leq x_j, \forall j; \sum_{j=1}^{n} d_{ij} \leq 1 \right\}. \tag{2}$$

The variable d_{ij} is a 0-1 variable where $d_{ij} = 1$ represents the admission of a type-j patient when a type-i patient is discharged. Clearly, this can happen only when there is at least one waiting

patient of type j in the queue. The constraint $d_{ij} \leq x_j$ forces d_{ij} to the value 0 when there is no waiting patient of type j. The constraint $\sum_{j=1}^{n} d_{ij} \leq 1$ states that we can admit at most one patient from all types. When all d_{ij} are zeros, it refers to choosing not to admit any patient.

3.3 Transition Probabilities

Let T denote the random time between two decision points. To find the distribution of T, we use Lemma 1 in EC.1. Based on the theorem in Porteus (2002), Lemma 1 establishes that the time to the next transition is exponentially distributed when all the events follow Poisson processes. The rate of the distribution is the sum of all rates; $\nu(\mathbf{x}, \mathbf{b}) = \sum_{i=1}^{n}(\lambda_i + b_i\mu_i)$. Also, when a transition has already happened at time t, the probability that the transition is caused by a specific event is the rate of that event divided by the sum of all rates. This probability is independent of the time that has passed. Since the state of the system changes over time, the transition rate in each period is not constant. To transform the system into a Markov chain with uniform transition rate, we apply the *uniformization* technique.

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

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

Now we can scale time such that the maximum transition rate ($\nu^{\max}$) is normalized to 1. To do so, we just need to define the new arrival and service rates: $\lambda_i{}' = \frac{\lambda_i}{\nu^{\max}}$ and $\mu_i{}' = \frac{\mu_i}{\nu^{\max}}$, for all i. Then the new transition probabilities are

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

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

3.4 The Bellman Optimality Equation

The Bellman equation of our dynamic program is given by

$$(\mathbf{DP}) \qquad h(\mathbf{x}, \mathbf{b}) = \boldsymbol{\pi}^T\mathbf{x} - \rho^* + \sum_{i=1}^{n} \lambda_i \min_{\mathbf{a}_i \in \mathcal{U}_i(\mathbf{x},\mathbf{b})} \{\kappa_i t_i + h(\mathbf{x} + (1 - a_i - t_i)\mathbf{e}_i, \mathbf{b} + a_i\mathbf{e}_i)\}$$

$$+\sum_{i=1}^{n} b_i\mu_i \min_{\mathbf{d}_i\in\mathcal{D}_i(\mathbf{x})}\{h(\mathbf{x}-\mathbf{d}_i,\mathbf{b}-\mathbf{e}_i+\mathbf{d}_i)\}$$
$$+\left(1-\sum_{i=1}^{n}\lambda_i-\sum_{i=1}^{n}b_i\mu_i\right)h(\mathbf{x},\mathbf{b}), \qquad \forall \mathbf{x},\mathbf{b},$$

where $\mathcal{U}_i(\mathbf{x},\mathbf{b})$ and $\mathcal{D}_i(\mathbf{x})$ are given by (1) and (2) respectively and $\mathbf{e}_i$ in an $n-$dimensional identity column vector.

In (**DP**), ρ^* is the optimal average cost per normalized time period and $h(\mathbf{x},\mathbf{b})$ is the bias function which represents the total difference from optimal average cost over all periods if we start from state $(\mathbf{x},\mathbf{b})$. The term $\boldsymbol{\pi}^T\mathbf{x}-\rho^*$ is the difference between the waiting cost of this period and the optimal average cost. The second (third) term refers to the case when a type-i patient arrives (discharges). The last term is associated with the case of no change in the state. This term has been added due to the uniformization. In our model, all the states can be reached from other states, i.e., *Weak Accessibility* holds (Bertsekas, 2005). Thus, the optimal average cost is independent of the initial state of the system.

4. The Data

In order to demonstrate the applicability of the proposed DP formulation and to garner managerial insights, we developed a full data set representing the patient flows through the 3-South neurology ward of Montreal Neurological Hospital (MNH). As mentioned before, the MNH neurology ward has sixteen inpatient beds. In an effort to focus on the care provided to stroke patients, we categorize the patients into four patient types: mild non-stroke, mild stroke, severe non-stroke, and severe stroke. Note that these patients arrive at the MNH through the ED and are kept boarding there until a bed becomes available at the ward. Our data set includes all patients treated in the neurology ward for three full fiscal years. We rely on three sources of data: the hospital's ED information system, the patient registry of McGill University Health Center, and the paper-based patient charts of the stroke and non-stoke patients admitted to the 3-South neurology ward of MNH.

In this section, we elaborate on the assumptions that we have made about the arrival and LOS distributions and verify them using the available data. Furthermore, we explain how the waiting costs in the model can be estimated in the form of health related quality of life (HRQoL) (Xie et al., 2006). We close the section by also describing how we estimate the cost associated with patient transfer, again in terms of HRQoL.

4.1 Analysis of Patient Arrivals

The patient inter-arrival times to the system are random and dependent on the type of the patient. We hypothesize Poisson distributions for the patient arrivals. The histograms of the number of

arrivals per day for each patient type based on three full years of actual data support this assumption. We test our hypothesis that the arrival process for each patient type follows a Poisson distribution using the χ^2 goodness-of-fitness test with bin size of one. The results are presented in Table 1, based on which we can conclude that Poisson distribution fits reasonably well to our data for each patient type. The goodness-of-fit χ^2 tests are found not statistically significant for all types, i.e., all p-values are >0.05.

Patient Type	Sample Size	Mean Number of Arrivals per Day	χ^2-test p-value
Mild Non-Stroke	259	0.236	0.097
Mild Stroke	289	0.262	0.165
Severe Non-Stroke	151	0.139	0.395
Severe Stroke	123	0.113	0.401

Table 1 χ^2 goodness-of-fitness tests for patient arrival process.

Hourly, daily, and monthly variations can play critical roles in modeling the arrival processes. To check whether the arrival rates in our problem vary with time of the day, day of the week, or month of the year we run Poisson regression analysis, using STATA 13, given that our data fits well to Poisson distribution. The Poisson regression model for each patient type uses 4380 points corresponding to the 6-hour time intervals in the three years from which the data has been collected. The relatively small arrival rates in our problem can result in many intervals with zero arriving patient. Also, the variance of arrival process tends be larger than the mean in this case. Therefore, we also conduct zero-inflated Poisson regression analysis. Comparing the zero-inflated Poisson regression models to the Poisson regression models using Vuong Non-Nested Hypothesis Test produces p-values of 0.099, 0.069, 0.240, and 0.178 for mild non-stroke, mild stroke, severe non-stroke, and severe stroke patients respectively. These p-values suggest that the zero-inflated Poisson regression models do not provide a significant improvement over the standard Poisson regression models. To confirm this, we also perform a log-likelihood ratio test to examine whether the zero-inflation component is in fact necessary. The results from this test present p-values greater than 0.05 for all patient types, which support the fact that the zero-inflated Poisson models are not significantly better than the Poisson models without zero-inflation component.

The results of the Poisson regression models, which are summarized in Tables 6–9 in EC.2, suggest that the number of arrivals do not vary with the time of the day, the day of the week, or

the month of the year for all patient types, given that all the p-values of corresponding categories of the variables are greater than 0.05. Our findings through zero-inflated Poisson regression models also report similar results, i.e. all p-values are greater than 0.05. In a further analysis, we define a binary variable of the weekend day instead of the day of the week variable and a variable of the season, which takes on values of fall (reference value), winter, spring and summer instead of the month of the year variable. The results of this analysis also confirm that the rates of arrivals do not vary either with the weekday/weekend or the season.

4.2 Analysis of Patient LOSs

The patient departures from the hospital during a given time period are random and dependent on the patient type. We hypothesize exponential service times for the LOS of patients. The histograms of LOS of each patient type support this assumption. We test our hypothesis that the LOS of each patient type follows an exponential distribution using Anderson-Darling goodness-of-fitness test. The Anderson-Darling statistic is not dependent on how the data is binned and does not require a sufficient sample size, which provides a more flexible method for our analysis. Using Minitab 17, we examine (i) if the data is from a population with exponential distribution and (ii) how well the data fits to other distributions such as lognormal, Weibull and Gamma. Table 2 summarizes the results from Anderson-Darling goodness-of-fitness test. Note that in fitting Weibull and Gamma distributions to our data we exclude the exponential distribution as a special form of these distributions.

Patient Type	Sample Size	Mean	Exponential		Weibull		Gamma		Lognormal	
			AD	p-value	AD	p-value	AD	p-value	AD	p-value
Mild Non-Stroke	259	13.003	1.129	0.084	1.062	<0.01	1.158	0.007	2.296	<0.005
Mild Stroke	289	11.491	1.229	0.064	1.328	<0.01	1.449	< 0.005	2.438	< 0.005
Severe Non-Stroke	151	19.011	0.716	0.263	0.744	0.05	0.794	0.048	0.810	0.035
Severe Stroke	123	22.002	0.428	0.589	0.421	>0.250	0.433	> 0.250	1.760	< 0.005

Table 2 Anderson-Darling (AD) goodness-of-fitness tests for LOS distributions.

Since a low p-value (<0.05) indicates that the LOSs do not follow that distribution, exponential distribution is the only one that fits to the data for mild stroke and non-stroke patients. For severe non-stroke patients, the data fits both to the exponential and Weibull distributions. However, given the lower AD and the higher p-value we choose the exponential distribution as the best fit. For severe stroke patients, although the lowest AD is for Weibull distribution, the AD values for

exponential, Gamma and Weibull are in the same range and results show a p-value higher than 0.05 for exponential distribution. Therefore, we conclude that exponential distribution also fits very well the LOS of severe stroke patients.

4.3 Patients Waiting and Transfer Costs

Waiting Cost: As mentioned earlier, the patients boarding in the ED for a bed in the neurology ward suffer from lack of specialized care and their health status deteriorates as a consequence of staying in the ED. This deterioration emerges as worse functionality of the patients, which is one of the most important health outcomes of the neurological patients. For those patients, discharge destination can be used as a proxy of patient's functionality at the time of discharge. In this context, Kucukyazici et al. (2010) have found that longer ED boarding time is strongly associated with increased probability of not being able to discharge to home, i.e., being admitted to rehabilitation center or long term care facility. To be more specific, they observed that 10% increase in the ED LOS is associated with 7.7% increase in the risk of being discharged to either a rehabilitation center or a long term care facility, i.e., not being able to go home. It is established that the discharge destination has a significant impact on both short-term and long-term HRQoL (Xie et al., 2006). Thus, we estimate the patient's waiting cost as the expected HRQoL lost resulting from not being able to go home due to the ED boarding.

Let β_i denote the percent increase in the probability of not being discharged home for type-i patients, as a result of one time unit of boarding in the ED. We estimate the patient type specific β_i utilizing a regression model controlled for all other clinical and demographic factors. Let s_i^{R} and s_i^{L} denote the conditional probabilities of being sent to a rehabilitation center and a long term care facility respectively given that the type-i patient is not discharged to home. Note that $s_i^{\mathrm{R}} + s_i^{\mathrm{L}} = 1$. Moreover, we define the HRQoL values associated with discharge destination as $\mathrm{Q_H}$, $\mathrm{Q_R}$, and $\mathrm{Q_L}$ for home, rehabilitation center and long term care facility, respectively. There are several studies in the literature that report the HRQoL measures for neurological patients including Hopman and Verner (2003), Jaracz and Kozubski (2003), Jönsson et al. (2005), and Nichols-Larsen et al. (2005). In our model, we use the short-term HRQoL measures estimated by Nichols-Larsen et al. (2005).

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

$$\pi_i = \beta_i p_i (\mathrm{Q_H} - (s_i^{\mathrm{R}} \mathrm{Q_R} + s_i^{\mathrm{L}} \mathrm{Q_L})) \tag{3}$$

where p_i corresponds to the average probability of discharge to rehabilitation center and long term care facility of patient type-i for the group of patients who do not experience any delay in the ED.

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

Table 3 Estimated patients' waiting cost in the ED per day

Using Equation (3), the waiting cost per day for each patient type is estimated and presented in Table 3.

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

5. Properties of the Optimal Policy: A Numerical Illustration

Before moving on to subsequent sections on approximation methods, we numerically explore the structure of the optimal admission policy by solving a large number of problem instances. By reporting the results from several revealing problem instances, we illustrate that the form of the optimal policy is not straightforward. The complexity of the problem stems from the fact that the optimal policy depends not only on how many beds are occupied, but also the number of beds occupied by each patient type, as well as the number of patients of each type waiting for a bed. It can be verified that the optimal policy is robust with respect to the magnitude of waiting and transfer costs and is affected only by their ratio.

We consider problem instances with two types of patients, mild stroke (referred to as type 1) and severe stroke (referred to as type 2). The arrival rates and average LOS for these two types are reported in Tables 1 and 2. Throughout this section, we assume the number of beds B and the

waiting room capacity K are both 8. We organize our illustrative examples into two subsections corresponding to the arrival and the discharge of a stroke patient, respectively.

5.1 Arrival of a Stroke Patient

Illustrative Example I: Let the waiting costs be $\boldsymbol{\pi} = (90, 295)$ and transfer costs $\boldsymbol{\kappa} = 2\boldsymbol{\pi}$. Suppose there is no severe stroke patient in the system; i.e., $x_2 = 0$ and $b_2 = 0$. When there is a new arrival of a mild stroke patient in the system, the optimal decision is to admit the new arrival to the ward if there is an available bed and transfer the patient otherwise. This implies that the last available bed is not reserved for a severe patient that may arrive in the future. However, if the waiting cost of severe stroke patients increases, the form of the optimal policy changes. In particular, when we increase π_2 to 450, the optimal policy is to reserve the last available bed for severe patients if the number of mild stroke patient waiting is no more than 5 ($x_1 \leq 5$). This example shows that a universal preference ordering between the two patient types does not exist.

Illustrative Example II: The condition $x_1 \leq 5$ in Example I implies a threshold policy to manage the last available bed. Denote the threshold on the number of mild patients waiting for a bed by γ, which equals 5 in Example I. In this example, we investigate the impact of the number of different patient types in the ward on γ. To this end, we consider situations where there is only one bed available (i.e., $b_1 + b_2 = B - 1$) and conduct a parametric analysis on the number of beds occupied by mild patients (b_1). We take $\boldsymbol{\pi} = (90, 450)$ and fix $x_2 = 0$. Figure 1 shows the optimal policy in the case of an arrival of mild stroke patient. When $b_1 \leq 2$, we reserve the last available bed for a severe patient if $x_1 \leq 3$ ($\gamma = 3$) and allocate that bed to the arriving mild stroke patient otherwise. This threshold increases by one (i.e., $\gamma = 4$) when $3 \leq b_1 \leq 4$, and increases by two (i.e., $\gamma = 5$) when $5 \leq b_1 \leq 7$. It is important to note that the states where $x_1 > 0$ are transient, since waiting mild patients are either transferred or admitted to beds. It is also interesting to observe that when there are less than four mild stroke patients waiting (as an initial state), the newly arriving mild stroke patients will be transferred regardless of the value of b_1. This example illustrates that the threshold on x_1 to manage the last available bed depends on the composition of patients that are already admitted.

Illustrative Example III: Here we consider a set of recurrent states in which $x_1 = x_2 = 0$. Let e be the number of empty beds. For all values of e, the optimal policy in the event of a mild patient arrival is shown in Figure 2. The cost parameters are the same as in Example II except that $\pi_2 = 325$. The optimal policy is to admit the arriving mild patient to the bed as long as more than two beds are available. If no bed is available, the patient is transferred. However, if only one bed is available, the admission policy depends on the patient mix in the ward. In particular, the arriving mild patient is admitted to the ward only if the number of beds occupied by mild patients

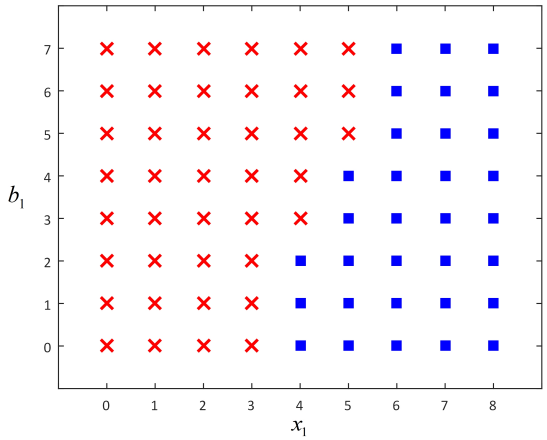

Figure 1 Optimal decision if a mild stroke patient arrives while $x_2 = 0$ and $b_1 + b_2 = 7$
× = Transfer (Reservation) ■ = Admit mild stroke patient to bed

(b_1) is greater than or equal to three. Therefore, in this case, the mild patient is admitted when *more* beds are already occupied by mild patients in the ward.

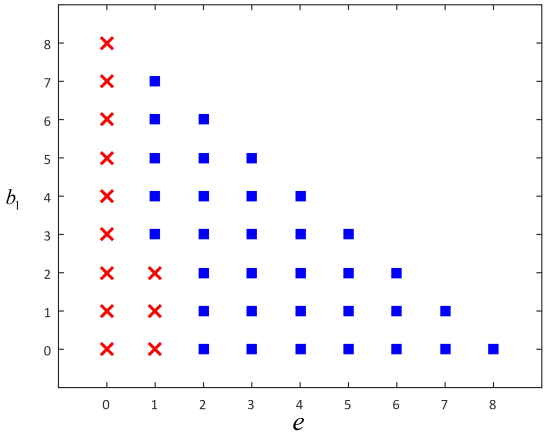

Figure 2 Optimal decision if a mild stroke patient arrives while $x_1 = x_2 = 0$, $b_1 + b_2 = B - e$
× = Transfer (Reservation) ■ = Admit mild stroke patient to bed

Illustrative Example IV: Now we consider the arrival of a severe stroke patient to the system. Similar to illustrative example II, we examine the optimal policy when $b_1 + b_2 = B - 1$ and $x_2 = 0$. When $\boldsymbol{\pi} = (90, 295)$, the optimal policy recommends that the newly arrived severe patient be admitted to the ward. However, when we decrease the waiting cost of severe patients from 295 to 135, it is optimal to transfer the arriving severe patients when $x_1 = K - 1$ and $b_1 \leq 1$ or $x_1 = K$ and $b_1 \leq 3$. This implies that the severe patients do not always get prioritized over the mild patients. Intuitively, when the system is highly congested and the chance of a discharge is low in the near

future, it is better to use the available beds to serve the mild patients while transferring the arriving severe patients.

5.2 Discharge of a Stroke Patient

Illustrative Example V: Consider the event of a patient discharge when all beds are occupied; i.e., one bed becomes available. The cost parameters are taken to be $\boldsymbol{\pi} = (200, 250)$ and $\boldsymbol{\kappa} = 2\boldsymbol{\pi}$. In Figure 3, we show the optimal decision when a mild stroke patient is discharged for all combinations of x_1 and x_2 and when $b_1 = b_2 = B/2$ before the patient discharge takes place. In this figure, observe that when $1 \leq x_1 \leq 3$ and one bed becomes available, we begin by assigning the emptied bed to a severe stroke patient. But if the number of severe stroke patients increases, it would be better to give that bed to a mild stroke patient. This seems to be counter-intuitive, as beds are assigned to mild patients even when there are severe patients waiting in the queue. The observation can be rationalized by considering the slower discharge rate of severe stroke patients. As the system becomes more congested, it becomes advantageous to serve the mild stroke patients who have higher discharge rate and hence a higher chance of emptying the beds in the near future. This phenomenon can happen when the waiting costs for the two types are close to each other. If we increase the waiting cost for severe patients, this phenomenon disappears. This is related to the observation in Example IV.

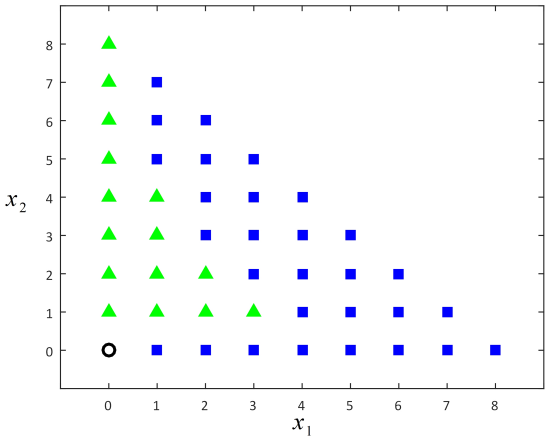

Figure 3 Optimal decision if a mild stroke patient is discharged while $b_1 = b_2 = 4$.
$\circ$ = No action ■ = Admit mild stroke patient to bed ▲ = Admit severe stroke patient to bed

Illustrative Example VI: In this example, we are interested in cases where it is optimal to reserve a newly freed bed for a future arrival of severe stroke patients. We take the cost parameters to be $\boldsymbol{\pi} = (100, 1500)$ and $\boldsymbol{\kappa} = 2\boldsymbol{\pi}$. We also assume $x_2 = 0$, otherwise bed reservation for severe

stroke patients cannot be optimal. When conduct a parametric analysis on b_1 when all beds are occupied (i.e., $b_1 + b_2 = B$). We vary b_1 from one to B to demonstrate how the reservation pattern changes depending on the patient mix in the ward in the event of a discharge of a mild stroke patient. Figure 4 depicts the optimal decision in this parametric analysis. Evidently, it is optimal to reserve a bed for severe stroke patients when their waiting cost is much higher than that of the mild stroke patients. Note that the threshold on x_1 above which we stop reserving increases as the number of occupied beds by mild stroke patients (b_1) increases.

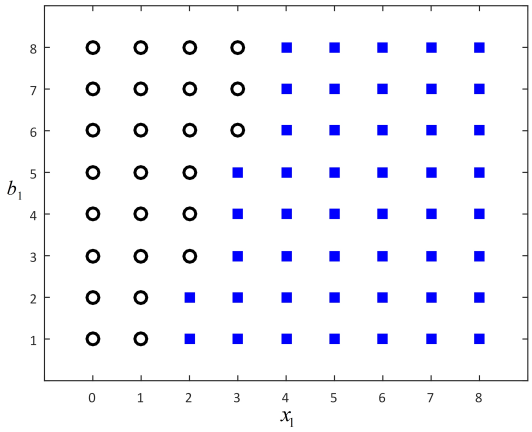

Figure 4 Optimal decision if a mild stroke patient is discharged while $x_2 = 0$ and $b_1 + b_2 = 8$.
$\circ$ = No action (Reservation) $\blacksquare$ = Admit mild stroke patient to bed

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

6. Solution Methodology

The Bellman equation for an average cost DP can be solved with the relative value iteration algorithm in a reasonable amount of time when the size of the problem is relatively small. As the number of patient types (n), the number of beds in the ward (B), or the waiting room capacity (K) increase, the *curse of dimensionality* hinders us from obtaining the optimal solution of the DP. Thus, we propose an approximation scheme to find a good admission policy in large-scale instances of the problem.

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

of patients from each type should be transferred to another hospital. The average cost of such a model accounts for waiting costs of patients as well as transfer costs. We call this model static because the policy is fixed over time irrespective of the state of the system.

The second step is to develop an ADP which can be solved in a reasonable time frame. To do so, we exploit some information from the static model's solution, including the opportunity cost of occupying a bed, the number of beds allocated to each type, the average waiting time, and the average queue length of each patient type. Then, the bias function $h(\mathbf{x}, \mathbf{b})$ is estimated using this information. To be more precise, we choose an arbitrary patient type. The value of bias function in state $(\mathbf{x}, \mathbf{b})$ is assumed to be the sum of some contributions from all patient types in that state. We estimate the contributions of all types of patients except that arbitrarily chosen one using the information extracted from the solution of the static model. For that specific type we leave the contribution unknown. We plug the estimated bias functions back into the Bellman equation to find the unknown contributions. This will lead to a DP with only one type of patient, which is simpler to solve. We iterate this procedure for all types of patient to find all the unknown contributions. At the end, we sum all the contributions up to approximate $h(\mathbf{x}, \mathbf{b})$. Based on the approximate bias function, we can create an admission policy. Figure 5 shows all these steps and their interactions.

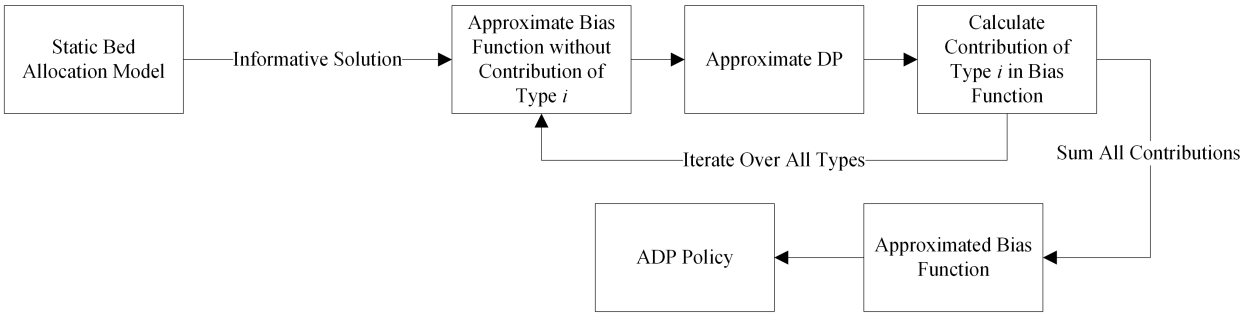

Figure 5 Schematic view of solution methodology

6.1 The Static Model

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

Suppose the number of beds dedicated to type-i patient is $\tilde{b}_i$. The system with $\tilde{b}_i$ beds serving incoming type-i patients can be viewed as a queue with $\tilde{b}_i$ servers. Due to the constraint on the total number of waiting patients, the type of queue we are dealing with for type-i patients is an

$M/M/\tilde{b}_i/\tilde{b}_i+k_i$ queue. Here k_i is the upper bound on the length of queue for type-i patients, above which the new arrivals will be turned away. The service rate is μ_i but the arrival rate should not necessarily be equal to λ_i, because we can transfer some patients upon their arrival to another hospital. So the rate of patients entering the system can be less than the original arrival rate. Therefore, we introduce a decision variable for the adjusted arrival rate as $\tilde{\lambda}_i$.

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

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

$$L=\begin{cases}\frac{p_{\rm o}r^c\rho}{c!(1-\rho)^2}[1-\rho^{k+1}-(1-\rho)(k+1)\rho^k], & (\rho\neq 1)\,,\\ \frac{p_{\rm o}r^c}{c!}\frac{k(k+1)}{2}, & (\rho=1)\,,\end{cases} \tag{4}$$

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

$$p_k=\frac{r^{c+k}}{c!c^k}p_{\rm o}, \tag{5}$$

where

$$p_{\rm o}=\begin{cases}\left[\frac{r^c}{c!}\left(\frac{1-\rho^{k+1}}{1-\rho}\right)+\sum_{n=0}^{c-1}\frac{r^n}{n!}\right]^{-1}, & (\rho\neq 1)\,,\\ \left[\frac{r^c}{c!}(k+1)+\sum_{n=0}^{c-1}\frac{r^n}{n!}\right]^{-1}, & (\rho=1)\,.\end{cases} \tag{6}$$

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

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

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

$$\sum_{i=1}^{n} k_i \leq K,$$
$$\tilde{\lambda}_i \leq \lambda_i, \quad \forall i,$$
$$\tilde{\lambda}_i, \tilde{b}_i \geq 0,$$
$$\tilde{b}_i, k_i \text{ integer}, \quad \forall i.$$

PROPOSITION 1. *The optimal solution of the (**SM**) gives an upper bound on the optimal average cost in the (**DP**); i.e., $F^* \geq \rho^*$.*

The proof of Proposition 1 is straightforward since the optimal solution of the static model (**SM**) is always a feasible policy for (**DP**).

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

1. If $\tilde{\lambda}_i = 0$, then $L_i = 0$ and $p_{k_i} = 0$.
2. If $\tilde{\lambda}_i \neq 0$ and $\tilde{b}_i = 0$, then $L_i = \infty$ and $p_{k_i} = 1$.
3. If $\tilde{\lambda}_i \neq 0$, $\tilde{b}_i \neq 0$ and $\tilde{b}_i$ is integer, then L_i and p_{k_i} are calculated through Equations (4) and (5).
4. If $\tilde{\lambda}_i \neq 0$ and $\tilde{b}_i \neq 0$ and $\tilde{b}_i$ is non-integer, then $\tilde{b}_i$ is rounded to nearest integer (called b_{new}) and service rate is adjusted to $\mu_{new} = \frac{\tilde{b}_i \mu_i}{b_{new}}$. The L_i and p_{k_i} are calculated using b_{new} and μ_{new}.

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

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

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

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

6.2 Approximate Dynamic Programming

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

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

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

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

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

$$
h(\mathbf{x},\mathbf{b}) \approx h_i(x_i, b_i) + \sum_{j \neq i} \left(\pi_j {W_j}^* (x_j - {L_j}^*)^+ + \frac{\alpha (b_j - \tilde{b}_j^*)^+}{\mu_j} \right), \quad \forall i, \tag{7}
$$

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

Similarly, the cost of occupying the bed by type-j patients is estimated. For each bed occupied in addition to the allocated beds in solution of (**SM**), $\tilde{b}_i^*$, the opportunity cost per unit time is

$\alpha(b_j - \tilde{b}_j^*)^+$. We know that on average, a typical patient of type j stays in bed for μ_j^{-1} units of time. Therefore, the total opportunity cost can be expressed by $\alpha\mu_j^{-1}(b_j - \tilde{b}_j^*)^+$.

Plugging (7) into (**LP**) and simplifying, we obtain a new linear program:

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

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

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

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

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

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

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

and

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

Note that $\mathcal{U}_k(\mathbf{x},\mathbf{b}) \subseteq \mathcal{U}_k'(x_i,b_i)$ and $\mathcal{D}_k(\mathbf{x},\mathbf{b}) \subseteq \mathcal{D}_k'(x_i)$. In the next step, we make the right hand side of the constraint dependent only on (x_i,b_i). Let us define $\mathbf{x}_{-i} = \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$ and $\mathbf{b}_{-i} = \{b_1, \ldots, b_{i-1}, b_{i+1}, \ldots, b_n\}$. Now, in order to make it independent of $\mathbf{x}_{-i}$ and $\mathbf{b}_{-i}$, we take the

minimum over $\mathbf{x}_{-i}$ and $\mathbf{b}_{-i}$ for each given (x_i, b_i). Consequently, by using the new action space and simplifying, the constraint of (**LP1**) will be a function of only x_i and b_i, and can be written as:

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

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

We can take out the minimization over $\mathbf{x}_{-i}$ and $\mathbf{b}_{-i}$ from above and write it as a separate mixed-integer program (**MIP**). We need to introduce binary variables to replace the indicator variables as well as some other binary and integer variables to remove the non-linear terms from the objective function. By taking all these steps, we will have an **MIP** with linear constraints and linear objective function, which is stated in detail in EC.3. The resulting **MIP** can be easily solved by CPLEX even with a large number of variables and constraints and we denote it by $\mathbf{MIP}(x_i, b_i, h_i(x_i,b_i))$ to emphasize its dependency on x_i, b_i and $h_i(x_i,b_i)$. By plugging back the **MIP** into the (**LP1**), we have:

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

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

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

PROPOSITION 2. *The optimal objective function of (**LP2**) gives a lower bound on the optimal average cost in (**DP**); i.e., $\rho_i^* \leq \rho^*$ for each i. Consequently, $\max_i \rho_i^* \leq \rho^*$.*

6.3 The ADP Policy

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

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

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

The ADP Policy:

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

7. Computational Experiments with Realistic Problem Instances

We consider problem instances with four patient types. Note that with four types of patients and a large number of beds, the optimal policy cannot be computed exactly due to the curse of dimensionality. In Section 7.1, we first describe the *Bed Allocation* policy and the *Bid Price* policy based on the solution of the static model (**SM**). We introduce six instances of the problem in Section 7.2, while a comparative analysis between the two static policies, i.e., the Bed Allocation and the Bid Price policies, the first-come-first-serve (FCFS) policy (as a benchmark), and the ADP policy over these six problem instances is reported in Section 7.3. Recognizing the difficulties

associated with the implementation of the ADP policy in practice, Section 7.4 presents the *Priority Cut-off* policy that is inspired by the ADP policy described in Section 6.3. In Section 7.5, we report on a second set of comparative analysis between the ADP policy, the ADP-based Priority Cut-off policy, and the current policy being used at the MNH. In Sections 7.6 and 7.7, we examine how the performance of the ADP policy is affected with respect to non-stationary patient arrivals and non-linear waiting cost functions.

7.1 Two Static Admission Policies

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

The Bed Allocation (BA) Policy:

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

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

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

The Bid Price (BP) Policy:

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

7.2 Six Problem Instances

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

We are unable to find the *optimal policy* for any of the six cases in our comparative studies. Nevertheless, it is possible to compare the ADP policy to the other heuristic policies. To this end, we developed a simulation model to help us find the average cost associated with a specific policy. The length of simulation horizon is considered to be 10,000 days with 1,000 days of warm-up period and we replicate the simulation for 100 times. Using the simulation results, we also report the average waiting time of all patients and the average transfer rates for each policy alternative. The simple averages do not reflect the true performance of each policy since transferring or ED

boarding a mild patient is not as undesirable as transferring one severe patient. Therefore, we use the unit time waiting costs (π_i) as weights to compute the weighted averages.

7.3 Comparative Analysis I

We now turn to a performance comparison among the First-come-first-serve (FCFS), Bed Allocation (BA), Bid Price (BP), and the ADP policies. The average total costs of the four admission policies for the six cases are depicted in Figure 6. In the figure, each plot corresponds to a case, and is located according to its congestion level (across the horizontal axis) and patient sensitivity to waiting (across the vertical axis). Evidently, the ADP policy produces the lowest average total cost in all cases. The other policy options fail to maintain low average total costs under all six patient sensitivity and congestion scenarios, e.g., the BP policy under case 3.

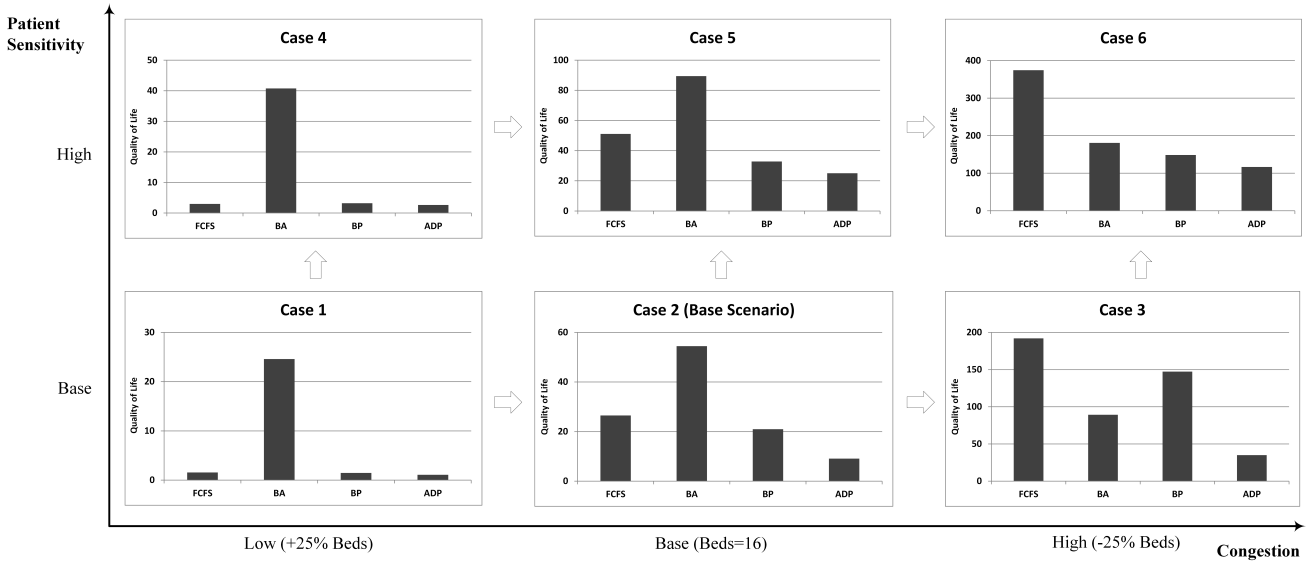

Figure 6 Average quality of life (QoL) lost per day – ADP policy versus static admission policies

The average waiting times and the average rates of patient transfer associated with the admission policy options are depicted in Figure 7. The trade off among these two performance measures is quite evident from this figure. The more a policy recommends transferring the patients to another hospital, the less the average waiting time experienced by the remaining patients. Note that the ADP policy seems to result in a more acceptable overall performance by balancing these two metrics. Even though the ADP policy does not produce the lowest average waiting time in all cases, its transfer rate is consistently reasonable.

In order to better display the comparison of the FCFS, BA, BP, and ADP policies, the plots in each of these two figures (and the two following figures) are not of the same vertical scale.

Consequently, these figures do not highlight the true impact of increased congestion and patient sensitivity levels on the three performance measures.

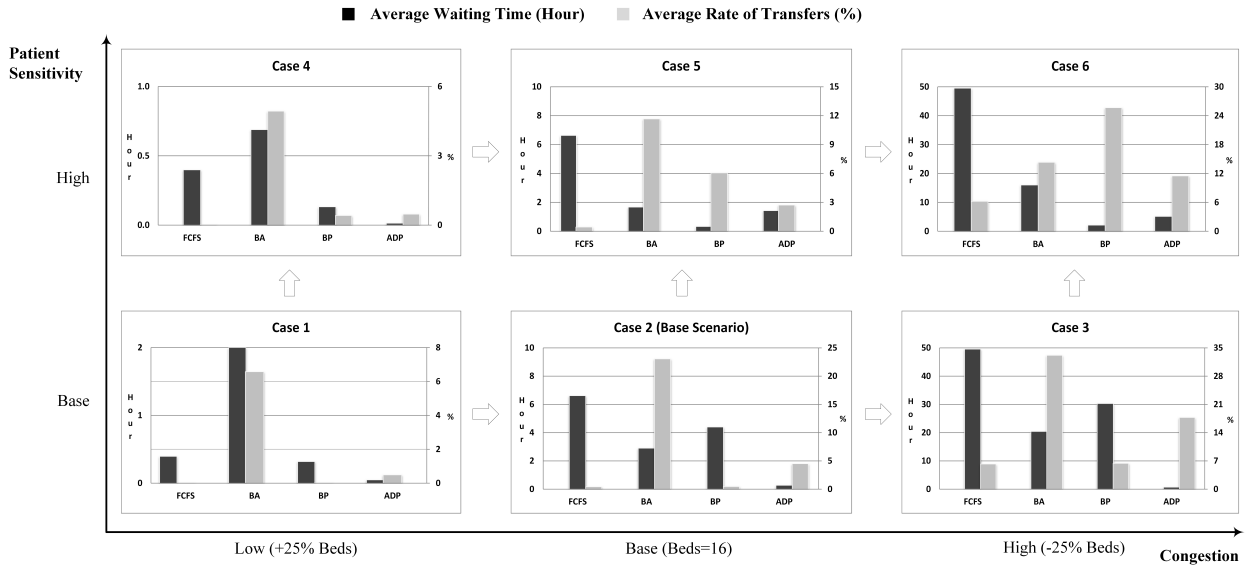

Figure 7 Average waiting time and rate of transfers – ADP policy versus static admission policies

We make the following observations:

1. When the transfer cost increases (i.e., moving up in Figure 7), all policies – except BP – decrease the rate of transfers, which results in longer waiting times.

2. When the system is more congested (i.e., moving right in Figure 7), the transfer rates increase in all policies in order to avoid much longer ED boarding times.

3. The BP policy in cases 1-3 is reduced to a simple priority queue with no transfers. This happens due to the small value of α and average waiting times obtained from the (**SM**). In all these cases, the BP policy also dominates the FCFS policy.

4. The BA policy does not seem to be very promising. The total average cost of this policy is almost the highest in all cases, except in case 3 where its transfer rate is not acceptable.

The overall managerial insight from Figures 6 and 7 is that, as the congestion and patient sensitivity levels increase, the ADP policy increasingly outperforms the other policies in terms of achieving both lower costs and acceptable trade offs between waiting times and patient transfers.

7.4 A More Practical Policy

The ADP policy can be challenging to implement as it provides an action for every state of the system. Through a detailed analysis of the results of the ADP policy, however, we observe that often only a few states of the system are critical in nature. For instance, when there is only one

bed available in the ward and a new patient arrives, the type of action we must take in response to the new arrival is crucial. Should we admit this new patient to the bed or save the last bed for the arrival of a more severe patient in the future? As a more general question, how many beds should we reserve for severe patients by not admitting the mild patients? Or, is reservation necessary at all? In contrast, making the best decision when half of the service capacity is available seems to be trivial. Thus, we develop a dynamic heuristic policy by following the ADP policy in the so-called critical states and applying a simple policy such as FCFS rule in other states, which would be much easier to implement.

In order to facilitate exploring the structure of the ADP policy, we first organize the patients into two groups regardless of their disease; mild and severe patients. The patients in the mild group have lower waiting cost and shorter average LOS; while in the severe group patients are highly sensitive to waiting and they occupy the bed for longer time periods. When we study the rules that constitute the ADP policy in six cases, it is evident that the severe patients should be prioritized over mild patients. The ADP policy always admits a severe patient if there is an available bed. However, this is not true for the mild patients. The ADP policy tends to reserve some beds for the severe patients (by not admitting the mild patients to those beds) unless there is a high chance of a patient discharge in the near future.

The chance of a future discharge depends on the patient mix in the ward, particularly the number of beds occupied by severe patients. Denote the aggregate number of severe patients staying in the ward by b_{s}. Based on the value of b_{s} at any time, we classify the chance of a discharge into three levels. The chance of a discharge is deemed high if $b_{\mathrm{s}} \leq \theta_1 B$, medium if $\theta_1 B < b_{\mathrm{s}} \leq \theta_2 B$, and low otherwise; $0 \leq \theta_1 < \theta_2 \leq 1$. We also introduce a threshold on the cost associated with transfer that affects the transfer decision. The transfer cost in this heuristic policy is defined to be small if $\boldsymbol{\kappa} \leq \omega \boldsymbol{\pi}$ and to be large, otherwise. The values for these thresholds can be derived based on the ADP policy recommendations at the critical states of the system.

Note that some simplifications are required to obtain the thresholds from the ADP policy. For example, in developing this heuristic policy we do not incorporate the number of patients in the queue in our admission decisions. This is justified by the results we obtained from the ADP policy in all six cases and it is mostly due to the low arrival rates of patients to the system in our examples. It is presumed that the queues are empty when a new patient arrives and consequently the decision is based only on the state of the ward. Therefore, the rules in this heuristic policy comply with the results of our illustrative example in Figure 2 of Section 5.

We call this heuristic policy the *ADP-based Priority Cut-off (PC)* policy because (i) it gives priority to certain types of patients, (ii) it changes behavior when the state of the system surpasses the cut-off points. Priority cut-off policies are commonly used in the context of patient scheduling

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

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

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

7.5 Comparative Analysis II

The second part of our analysis in this section involves comparing the ADP-Based Priority Cut-off (PC) policy and the current policy being used in the MNH with the ADP policy. The MNH policy has been briefly discussed in Section 1. It allocates a fixed number of beds to each patient type regardless of their level of severity and leave some beds flexible to be used by all patient types. To be more specific, six beds out of 16 available beds are dedicated to stroke patients, same number of beds are allocated to non-stroke patients, and the rest of beds are being used by both type of patients.

Let us denote the number of beds dedicated to stroke patient beds by b_{stroke}, the number of beds dedicated to non-stroke patient beds by $b_{\text{non-stroke}}$, and the number of flexible beds by b_{flexible}. The patients are admitted to the beds until all the dedicated beds to their type and flexible beds are full. Then, they wait in the queue for a bed in the ward until the waiting time exceeds a threshold (denoted by d) in which case they have to be transferred. The hospital uses the same time threshold for all patient transfers. We summarize this policy, which is a static bed allocation policy, below.

The Current (MNH) Policy:

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

For the base case (case 2) we know that $(b_{\text{stroke}}, b_{\text{non-stroke}}, b_{\text{flexible}}) = (6, 6, 4)$. For other cases, we adjust the bed allocations by simply keeping the same ratios as in case 2 between the beds assigned to different patient types. To find the transfer threshold (d), we follow the same idea we used in section 4.3 to estimate $\boldsymbol{\kappa}$. As we have set $\boldsymbol{\kappa} = 3\boldsymbol{\pi}$ for cases 4-6, this implies the transfer threshold for the patients in these cases is three days (72 hours). Therefore, the current (MNH) policy uses the following parameters in our experiments:

- In cases 1 and 4, we have $(b_{\text{stroke}}, b_{\text{non-stroke}}, b_{\text{flexible}}) = (8, 8, 4)$.
- In cases 2 and 5, we have $(b_{\text{stroke}}, b_{\text{non-stroke}}, b_{\text{flexible}}) = (6, 6, 4)$.
- In cases 3 and 6, we have $(b_{\text{stroke}}, b_{\text{non-stroke}}, b_{\text{flexible}}) = (5, 5, 2)$.
- In cases 1-3, we have $d = 48$ hr, and In cases 4-6, we have $d = 72$ hr.

Here, we also remind the reader that the thresholds of the ADP-based PC policy explained in Section 7.4 vary with the cost parameters. By examining the results of the ADP policy for the six cases, we observed the following:

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

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

The average total costs of the ADP, PC and MNH policies are depicted in Figure 8. The ADP policy has the lowest average cost in all cases, whereas the costs associated with the PC policy are consistently within an acceptable range of the ADP policy. The difference between the ADP policy (or PC policy) and the MNH policy is more pronounced when the patients are more sensitive to waiting and service capacity is limited (i.e., cases 2-3 and 5-6).

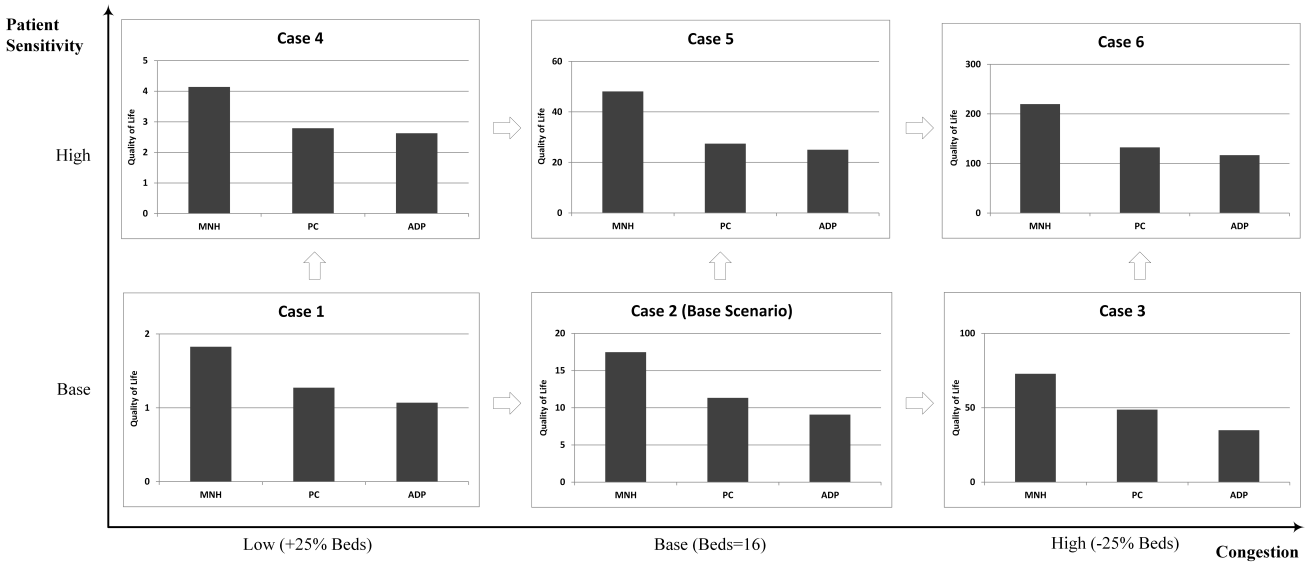

Figure 8 Average quality of life (QoL) lost per day – ADP policy versus practical admission policies

The average waiting time and average rates of patient transfer for the three policies are shown in Figure 9. The PC policy is more conservative than the ADP policy in terms of patient transfers. In all cases, it transfers fewer patients and consequently it has higher average waiting times. Compared to the current policy, the ADP policy decreases the waiting time significantly while its transfer rates are slightly higher in some cases. The PC policy, however, reduces the wait times in most cases by transferring the same or less number of patients. Hence, it could be utilized as an efficient and practical policy by the hospital to improve the performance of the ward in terms patients' health outcomes. It is also important to note that the PC policy generates the second lowest average costs over six cases compared to the static policy alternatives (i.e., BA, BP, and FCFS policies) in Section 7.3.

7.6 Non-Stationary Arrivals

Our modeling framework and solution approach assume stationary arrival process. The stationarity assumption is customary in the healthcare operations literature (Patrick et al., 2008) and is verified using the data obtained from MNH in Section 4. Nevertheless, we show that our work can

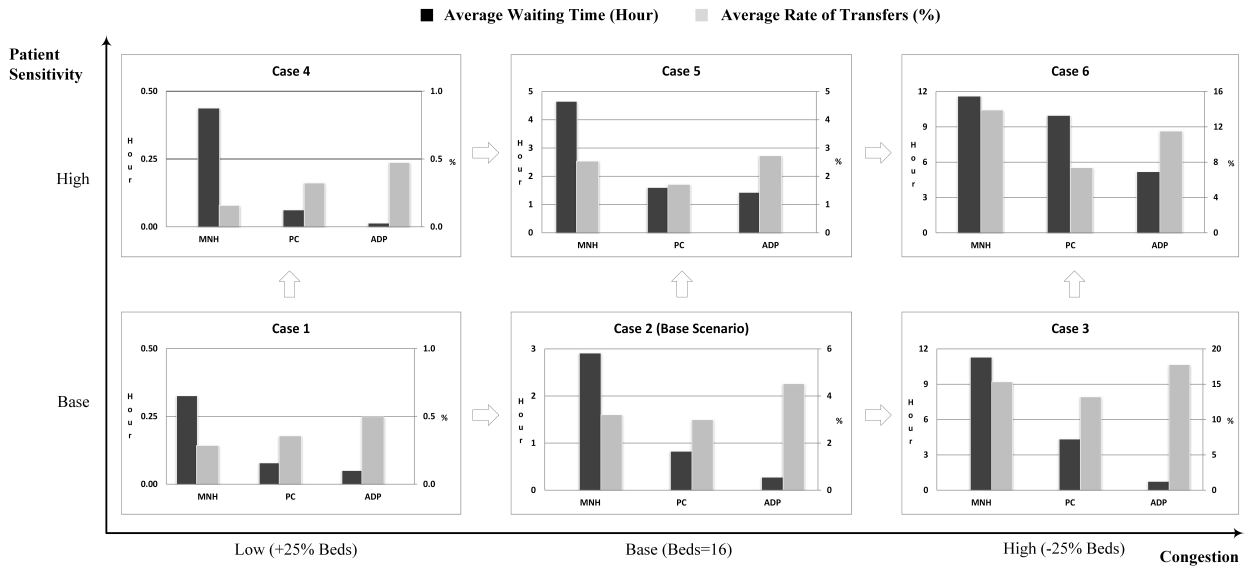

Figure 9 Average waiting time and rate of transfers – ADP policy versus practical admission policies

be adapted to problem settings with non-stationary arrival processes by building on point-wise stationary approximation (PSA) in the queueing control literature (Green and Kolesar (1991) and Yoon and Lewis (2004)). We conduct numerical experiments with non-stationary arrival processes to check the robustness of our proposed ADP approach.

The PSA approach combines solutions from problems with stationary arrivals to construct heuristic control policies for problems with non-stationary arrival processes. For each problem with stationary arrivals, the ADP approach proposed before can be applied to obtain a solution. Our numerical instances consider problems with periodically time-varying arrival rates; that is, the patient arrival rates follow a cyclic pattern that repeats itself after a given period of time. This type of time-varying arrival rates are considered in both Green and Kolesar (1991) and Yoon and Lewis (2004).

Assume that there are τ points at which the arrival rates of patients change (i.e., there are $\tau+1$ different arrival rates within the cycle). Then each cycle involves $\tau+1$ sub-period with stationary arrival rates. To obtain a solution via PSA, a separate problem is solved for each sub-period where we assume the prevailing arrival rate in the subperiod remains constant over the infinite time horizon. The resulting problems have stationary parameters, which are then solved using the solution approach we introduced earlier in the paper. Each stationary problem solved in this fashion suggests an admission policy. To construct a heuristic policy for the original problem with non-stationary arrival rates, we piece together policies from the stationary problems, where the stationary policy from each problem is only implemented for the relevant intervals with constant arrival rates. Note that the heuristic approach introduced here can be applied for the ADP, PC, as well as the BA and BP policies.

To examine the performance of PSA policies, we consider problem instances where the arrival rates follow a weekly cyclic pattern. Patients arrive at the hospital at higher rates during the first three days of a week, i.e., between Mondays and Wednesdays. As we move towards the end of a week, the arrival rates tend to decrease. Specifically, on Thursdays and Fridays, patient arrival rates are reduced by 50%. The arrival rates are again halved during weekends (Saturdays and Sundays). Note that there are two points during a week at which the arrival rates change, i.e., $\tau = 2$. Therefore, to solve the problem using the PSA approach, we need to solve three $(\tau + 1)$ problems with stationary arrival rates. We have chosen the time-varying arrival rates such that the arrival rates on Thursdays and Fridays are equivalent to the values for Cases 1-6 in Section 7.2. The LOS distributions and other problem parameters are also the same as Cases 1-6 in Section 7.2.

We use PSA to construct all policies except FCFS. Table 4 reports average QoL lost per day for the five policies. The last column of Table 4 shows the percentage improvement of the ADP policy, compared with the best alternative policies. For Cases 1-3, the best alternative policy is PC, while the best alternative policy for Cases 4-6 is BP.

To compare the ADP policy to other alternatives, we find the non-stationary BA and BP policies by solving the **SM** model with arrival rates of each period. The PC policy again is developed based on the ADP policy. Both BP and PC are significantly better than FCFS policy in all cases. Overall, the ADP policy has the best performance, although it is slightly inferior to BP in Cases 5-6, both with very high system congestion level. Therefore, the performance of the ADP policy seems to deteriorate as the system is heavily congested. For all other cases, the ADP policy shows significant improvement over alternative policies. Overall, our numerical results suggest that the proposed ADP policy continues to be a reasonable solution approach for problem instances with non-stationary problem parameters, which is consistent with the observations in the PSA literature we cited earlier.

Case	Policy					ADP
	FCFS	BA	BP	PC	ADP	Improvement
1	2.84	36.36	2.42	1.81	1.59	13%
2	29.37	60.24	24.22	12.06	9.99	17%
3	170.70	89.50	83.61	64.83	63.26	2%
4	5.44	341.20	4.60	5.09	3.92	15%
5	56.59	116.69	54.93	57.12	55.74	-1%
6	333.61	151.96	140.85	146.33	140.99	-0.1%

Table 4 Robustness of the ADP policy with respect to non-stationary arrivals

7.7 Non-linear Cost Functions

In this section, we consider non-linear waiting cost functions as it seems more realistic to assume the patient's health status deteriorates at higher rates when the waiting time increases. To this end, we speculate convex piecewise-linear increasing functions for the waiting costs of patients. For each patient type, we divide the time between zero and the transfer threshold into three-hour intervals. During each interval the waiting cost is a linear function of time with a slope that is increasing from one interval to another. The slope increases until the transfer threshold, after which it remains constant and equal to slope of the last interval.

To find the patient admission policy using the proposed ADP approach, we need to linearize the cost functions. We run regression models with zero intercept to fit linear lines to the non-linear cost functions. Using the slopes of the fitted linear functions for each patient type, we solve the ADP and obtain the associated policy. In order to compare the performance of the admission policies under linear and non-linear waiting cost structure, we choose the parameters of the non-linear cost function such that the slope of the fitted line is equal to the waiting cost per unit time (π_i) considered in the problem instances of Section 7.2. We then use simulation to calculate the average total cost in both scenarios.

The results of this comparison are shown in Table 5. In this table, the percent increase in the total cost associated with each policy, when the waiting costs are incurred according to a non-linear function is reported. The last two columns of this table show the percent improvement achieved by the ADP policy in each scenario over the best of the other policy options. A negative percentage implies that the ADP policy is dominated by another heuristic policy. From the table, it can be concluded that the performance of the ADP policy remains robust to the change in the structure of waiting costs in almost all cases. Except in case 4, in which there is enough service capacity, the percent improvement of the ADP policy over other policies has in fact increased.

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

Table 5 Robustness of the ADP policy with respect to non-linearity of waiting cost

8. Concluding Remarks

We have considered an admission control and bed allocation problem, incorporating the differentiating features of neurology wards. From a modeling perspective, we presented an average cost DP that assumes none of the beds in the ward are earmarked to certain patient types. To overcome the *Curse of Dimensionality* of DP formulation that prevents us from solving the realistic-size problem instances in reasonable amount of time, we proposed an ADP that uses some information from a static queuing model. To the best of our knowledge, the ADP for the average cost problem has not been fully explored theoretically. A couple of examples are Roubos and Bhulai (2010) and Roubos and Bhulai (2012) that use ADP in controlling queues with application to call centers. The numerical results from our experiments on some problem instances, based on Montreal Neurological Hospital, revealed that the admission policy suggested by our proposed ADP works very well compared to the other heuristic policies we have studied.

Recognizing the managerial challenges in implementing the fully state dependent ADP policy, we also developed an ADP-based Priority Cut-off policy that performs quite well. We must emphasize that the structure of this heuristic policy is highly dependent on the results of our experiments for the six problem instances we have considered in the comparative analysis. The goal of developing such a policy was to demonstrate how an easy-to-use set of admission rules can be derived from the ADP policy for hospital managers.

The current admission policy at the hospital involves dedicating six beds to stroke patients and six beds to non-stroke patients, while leaving four beds flexible for both patient types. Furthermore, a patient transfer request is triggered after 48 hours of ED boarding. In contrast, the proposed ADP policy does not use earmarked beds and decides to transfer the patient at the time of arrival, considering the state of the system. By comparing these two policies, we observe that the current policy can be 70-110% worse than the ADP policy in terms of average HRQoL lost per day. Also, the ADP policy can decrease the average boarding time in the ED (especially when there are limited number of beds available such as in case 3 and case 6 of our comparative analysis) significantly without affecting the average rate of patient transfers. Thus, we provide the following insights for neurology ward managers: (i) it is better to decide whether or not to transfer a patient to another hospital immediately upon arrival and by taking into account the state of the system, (ii) dedicating neurology ward beds to patient types can worsen average ED boarding times, (iii) if the managers prefer to use an earmarking strategy, it is recommended to do so based on the level of severity of the patients condition rather than their disease (i.e., along the lines of the PC policy).

The modeling framework proposed in this paper is based on two simplifications. First, a small percentage of the patients with neurological conditions can be admitted directly to the ward for

elective surgeries, while this paper is confined to the patients who are admitted through the ED. Second, some patients, e.g., severe stroke patients, require intensive care for stabilization prior to being admitted to the ward, which we do not represent in our model. Their LOS in the Neuro-ICU, however, is most often 48 hours with fairly low variability. Extensions to the model proposed in this paper to relax these two assumptions constitute fruitful avenues for future research.

The primary structural assumption underlying the model presented in this paper is the medical infeasibility of caring for neurology patients at off-service beds, i.e., beds located in other wards of the hospital. Although the use of off-service beds has been common practice in the health sector, it constitutes a short-term fix for, arguably, prevailing systemic issues in the institution. This is analogous to the use of inventories to offset the underlying capacity imbalance between sequential manufacturing processes. The impact of an off-service hospital admission on the health outcome varies among patient types. According to Singh et al. (2012), for example, the admission of oncology patients in hallway or in off-service beds did not appear to compromise the timeliness or frequency of medical assessments. However, delays in nursing care were observed and patient satisfaction was decreased. For acute heart failure (AHF) patients, however, the negative impact of an off-service admission can be serious. Cowie et al. (2013) pointed out that mortality can be reduced if AHF patients are rapidly and accurately assessed in the ED, and admitted to a cardiology ward with the required expertise. According to the National Heart Failure Audit conducted in 2011-12, only about half of AHF patients were treated in cardiac wards. The cardiac ward had 7.8% mortality rate for AHF, whereas mortality was significantly higher at the general medicine and other wards (13.2% and 17.4%, respectively). Evidence also suggests that the mortality advantage for cardiology ward treatment persists post-discharge.

We witness a current trend among hospital managers to minimize the use of off-service beds as part of their efforts to improve the patient flow through the hospital. A recent effort at the Rouge Valley Health System (RVHS) for the development and implementation of a new bed map (Williams and Topaloglou, 2013) is a good example. RVHS is a two-site hospital with 479 beds serving the East Greater Toronto Area. One of the main motivations for the process redesign was the "significant problems with off-servicing". RVHS was able to reduce the medical off-service beds from 20 to one by the new bed map. Another example is Vibra Hospital of the Central Dakotas. This is a specialty acute care hospital that provides long-term acute care to complex patients with multiple comorbidities, requiring an extended stay in a hospital setting. Multidisciplinary teams comprising up to eight individuals provide specialty care, and hence it is not medically acceptable to admit a patient to an off-service area in Vibra. The model proposed in this paper would be applicable to the extent that off-servicing in the hospital is significantly reduced as an administrative policy.

On the methodological side, we contribute to the literature on approximate dynamic programming (ADP) for admission control of queues. Allocation of a limited capacity of resources among several customer types is a critical decision faced in many settings, including healthcare (Gupta, 2013), telecommunications (Paschalidis and Tsitsiklis, 2000), and manufacturing (Buzacott and Shanthikumar, 1993). While it is common to formulate queueing control problems in the dynamic programming framework, solving the resulting problems exactly is usually not feasible due to the large state and/or control spaces. The approach proposed in our paper takes two steps. First, a queueing control problem is formulated under a *static* control policy. Second, solution from the first step is used to build value function approximations in the linear programming based ADP framework (see, e.g., de Farias and Van Roy, 2003). Our approach is applicable when i) queueing control under a static policy is tractable, and ii) the approximate linear program resulting from the value function approximation can be solved. While our paper gives one such example, there are potentially many other problems where the approach is applicable.

Specific to the healthcare operations context, dynamic programming models of queueing admission control have been very popular (see, e.g., Ayvaz and Huh, 2010, Helm et al., 2011, Green et al., 2006). In the aforementioned papers, heuristic control policies are usually based on the analysis of special cases with certain parameter restrictions on the queues (e.g., all customer types share the same service rate). In contrast, we build dynamic heuristic control policies based on the analysis of static control policies; no parameter restrictions are imposed on the queues. While it is certainly outside the scope of the current research, it is an interesting future research topic to compare and contrast the two approaches. One advantage of the approach proposed in our work is its sound theoretical foundation in the general framework of linear programming based ADP.

Within the domain of linear programming based ADP, value function approximations are generally linear and/or separable. A somewhat unique aspect of our work is that the value function approximation is nonlinear and nonseparable. We show that the approximate linear program resulting from such a value function approximation is still tractable. We certainly hope that more researchers will adopt such approximation architecture in their work in the future.

Acknowledgments The authors are grateful to Dr. Richard Riopelle, Chair of Neuro-surgery at the Montreal Neurological Hospital, for his insights and providing us with the access to the hospital data collection. This research is supported in part by Natural Sciences and Engineering Research Council of Canada through Discovery Grants to the third and fourth authors. We acknowledge the constructive criticism from the two referees, the Associate Editor and the Area Editor, which were helpful in improving the manuscript.

Saied Samiedaluie is an assistant professor in the Department of Accounting, Operations and Information Systems at the Alberta School of Business, University of Alberta. His research interests

focus on developing data-driven analytical models to support decision making in health care and medicine, in particular decisions related to health care operations management, clinical decision making, and health policy design. He is interested in exploring innovative combinations of data sciences and operations research methodologies to study complex, practically relevant problems in these areas.

Beste Kucukyazici is an assistant professor of Operation Management at the Desautels Faculty of Management at McGill University. She is also a Research Associate at the St. Mary's Research Center. Dr. Kucukyazici's research interest is in decision-making problems under uncertainty with special interest in the health care policy design, health care operations, optimization of medical interventions and medical decision-making. She is particularly interested in the mathematical modeling and analysis of such problems through the use of the methodologies of Markov decision processes, stochastic analytical models, simulation and biostatistics.

Vedat Verter is James McGill Professor of Operations Management at the Desautels Faculty of Management of McGill University. He specializes on the application of business analytics for policy design and decision-making in the public sector. His areas of research are transport risk management, sustainable operations and healthcare operations management. His work in these areas is well recognized through top tier journal publications as well as invited presentations around the globe. In the area of healthcare, he focuses on preventive, primary, emergency, acute and chronic care processes, as well as their interaction. He is Founding Director of the NSERC CREATE Program in Healthcare Operations and Information Management, a seven-University PhD/PDF training program across Canada. Professor Verter is also Editor-in-Chief of Socio-Economic Planning Sciences, an international journal focusing on public sector decision-making.

Dan Zhang is an associate professor in the Management and Entrepreneurship division at Leeds School of Business, University of Colorado Boulder. Dr. Zhang received his PhD in industrial engineering from University of Minnesota and subsequently did postdoctoral work at Booth School of Business, University of Chicago. In the last few years, he worked on approximate dynamic programming methods for large-scale dynamic optimization problems and on consumer behavior models with applications to revenue management and pricing.

References

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

Bertsekas DP (2005) *Dynamic programming and optimal control* (Athena Scientific).

Buzacott JA, Shanthikumar JG (1993) *Stochastic models of manufacturing systems*, volume 4 (Prentice Hall Englewood Cliffs, NJ).

Carr S, Duenyas I (2000) Optimal admission control and sequencing in a make-to-stock/make-to-order production system. *Operations research* 48(5):709–720.

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

Chalfin DB, Trzeciak S, Baumann ALBM, Dellinger RP (2007) Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Critical Care Medicine* 35(6):1477–1483.

Collaboration SUT (2013) Organised inpatient (stroke unit) care for stroke. *Cochrane Database Syst Rev* 9.

Cowie M, Bell D, Butler J, Dargie H, Gray A, McDonagh T, McIntyre H, Squire I, Taylor J, Williams H (2013) Acute heart failure a call to action. *British Journal of Cardiology* 20(suppl 2):S1–S11.

de Farias DP, Van Roy B (2003) The linear programming approach to approximate dynamic programming. *Operations Research* 51(6):850–865.

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

De Vericourt F, Karaesmen F, Dallery Y (2002) Optimal stock allocation for a capacitated supply system. *Management Science* 48(11):1486–1501.

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

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

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

Gross D, Shortle JF, Thompson JM, Hartis CM (2008) *Fundamentals of queueing theory* (John Wiley and Sons).

Gupta D (2013) Queueing models for healthcare operations. *Handbook of Healthcare Operations Management*, 19–44 (Springer).

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

Hopman WM, Verner J (2003) Quality of life during and after inpatient stroke rehabilitation. *Stroke* 34(3):801–805.

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

Jönsson AC, Lindgren I, Hallström B, Norrving B, Lindgren A (2005) Determinants of quality of life in stroke survivors and their informal caregivers. *Stroke* 36(4):803–808.

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

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

Kucukyazici B, Green L, Verter V (2010) Improving stroke outcomes through operational policies. Extended abstract, Annual Conference of MSOM, Haifa, Israel.

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

Li X, Beullens P, Jones D, Tamiz M (2008) An integrated queuing and multi-objective bed allocation model with application to a hospital in china. *Journal of the Operational Research Society* 60(3):330–338.

Lopez AD, Mathers CD, Ezzati M, Jamison DT, Murray CJ (2006) *Global burden of disease and risk factors* (Oxford University Press, USA).

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

Nichols-Larsen DS, Clark P, Zeringue A, Greenspan A, Blanton S (2005) Factors influencing stroke survivors quality of life during subacute recovery. *Stroke* 36(7):1480–1484.

Paschalidis I, Tsitsiklis J (2000) Congestion-dependent pricing of network services. *Networking, IEEE/ACM Transactions on* 8(2):171–184.

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

Porteus EL (2002) *Foundations of Stochastic Inventory Theory* (Stanford University Press).

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

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

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

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

Singh S, Trudeau ME, Imrie KR, De Mendonca B, Fralick J, Cheung MC (2012) The association between the transfer of emergency department boarders to inpatient hallways or off-service beds and the quality of oncology care. *ASCO Annual Meeting Proceedings*, volume 30, 113.

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

Williams C, Topaloglou T (2013) Implementing a new bed map at rouge valley health system: A collaborative approach to aligning care delivery and patient need. Technical report, Rouge Valley Health System, URL `http://www.nhlc-cnls.ca/assets/7_Implementinganewbedmap_Williams.pdf`.

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

Xie J, Wu EQ, Zheng ZJ, Croft JB, Greenlund KJ, Mensah GA, Labarthe DR (2006) Impact of stroke on health-related quality of life in the noninstitutionalized population in the united states. *Stroke* 37(10):2567–2572.

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

Appendices

EC.1. Transitions Probabilities

Here, we present the lemma by Porteus (2002) that is used to calculate the transition rates and probabilities.

Lemma EC.1.

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

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

EC.2. Analysis of Patient Arrival Stationarity

To investigate whether the arrival rates vary over time we run Poisson regression analysis, using STATA 13, given that our data of arrivals fits very well to Poisson distribution. The Poisson regression model for each patient type uses 4380 points, corresponding to 6-hour time intervals over three year. In the regression models, we define the number of arrivals within 6 hour time interval as a dependent variable where we have three independent variable: time of the day, day of the week and month of the year. Time of the day is defined as a categorical variable which takes on the values of 12 a.m.–6 a.m. (reference value), 6 a.m.–12 p.m., 12 p.m.–6 p.m., and 6 p.m.–12 a.m. The variable of day of the week takes on the values of Monday to Sunday, where Monday is the reference value. In a similar way, the variable of month of the year takes on the values from January to December, where January is the reference value. The results of the Poisson regression models are shown in Tables EC.1-EC.4.

EC.3. The Mixed-Integer Program

In this section, we provide the details of the mixed-integer program that was used in the last step of ADP. We pull out the minimization problem over $\mathbf{x}_{-i}$ and $\mathbf{b}_{-i}$ as a separate **MIP** and solve it in every iteration of the ADP. To solve the **MIP** efficiently we need to define some binary variables to replace the indicator variables in the minimization problem. The first set of variables is:

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

Patient Type		Estimate	Std. Error	Z-Value	P-Value	95% CI
Mild Non-Stroke	**Month of the Year**					
	February	-0.05	0.31	-0.15	0.88	(-0.66,0.57)
	March	-0.10	0.31	-0.32	0.75	(-0.70,0.51)
	April	-0.35	0.34	-1.05	0.30	(-1.01,0.31)
	May	0.40	0.28	1.47	0.14	(-0.14,0.95)
	June	-0.07	0.31	-0.21	0.83	(-0.67,0.54)
	July	-0.10	0.31	-0.31	0.76	(-0.70,0.51)
	August	-0.32	0.33	-0.97	0.33	(-0.96,0.33)
	September	-0.02	0.31	-0.05	0.96	(-0.61,0.58)
	October	0.00	0.30	-0.10	0.99	(-0.59,0.59)
	November	0.08	0.30	0.26	0.80	(-0.51,0.66)
	December	0.24	0.29	0.84	0.40	(-0.32,0.80)
	Day of the Week					
	Tuesday	-0.14	0.24	-0.60	0.55	(-0.62,0.33)
	Wednesday	-0.03	0.23	-0.11	0.91	(-0.49,0.43)
	Thursday	-0.14	0.22	-0.66	0.51	(-0.29,0.59)
	Friday	0.08	0.23	0.36	0.72	(-0.36,0.53)
	Saturday	0.06	0.23	0.25	0.80	(-0.39,0.51)
	Sunday	-0.14	0.24	-0.59	0.56	(-0.62,0.33)
	Hour of the Day					
	6 a.m.–12 p.m.	-0.01	0.17	-0.07	0.95	(-0.35,0.33)
	12 p.m.–6 p.m.	-0.10	0.18	-0.57	0.57	(-0.44,0.25)
	6 p.m.–12 a.m.	-0.09	0.18	-0.51	0.61	(-0.43,0.26)
	Constant	-2.78	0.28	-9.89	0.00	(-3.33,-2.23)

Table EC.1 Poisson regression analysis for estimating number of arrivals of mild non-stroke patients

along with following constraints:

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

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

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

Patient Type		Estimate	Std. Error	Z-Value	P-Value	95% CI
Mild Stroke	**Month of the Year**					
	February	0.27	0.29	0.92	0.36	(-0.30,0.84)
	March	0.27	0.28	0.97	0.33	(-0.28,0.83)
	April	0.24	0.29	0.82	0.41	(-0.28,0.83)
	May	0.37	0.28	0.82	0.41	(-0.33,0.80)
	June	0.16	0.29	0.54	0.59	(-0.42,0.73)
	July	-0.10	0.31	-0.31	0.76	(-0.70,0.51)
	August	0.00	0.30	0.00	0.99	(-0.59,0.59)
	September	0.20	0.29	0.69	0.49	(-0.37,0.77)
	October	0.13	0.29	0.43	0.67	(-0.45,0.70)
	November	0.38	0.30	1.35	0.18	(-0.17,0.92)
	December	0.54	0.27	1.93	0.05	(0.02,1.07)
	Day of the Week					
	Tuesday	0.00	0.21	0.00	0.99	(-0.42,0.42)
	Wednesday	0.17	0.21	0.81	0.42	(-0.24,0.57)
	Thursday	0.04	0.21	0.19	0.85	(-0.37,0.45)
	Friday	0.15	0.21	0.73	0.46	(-0.25,0.55)
	Saturday	0.07	0.21	0.35	0.72	(-0.34,0.49)
	Sunday	-0.12	0.22	-0.53	0.60	(-0.55,0.32)
	Hour of the Day					
	6 a.m.–12 p.m.	-0.04	0.16	-0.23	0.82	(-0.35,0.28)
	12 p.m.–6 p.m.	0.07	0.15	0.45	0.65	(-0.23,0.37)
	6 p.m.–12 a.m.	-0.05	0.05	-0.33	0.74	(-0.37,0.26)
	Constant	-2.78	0.28	-10.56	0.00	(-3.41,-2.34)

Table EC.2 Poisson regression analysis for estimating number of arrivals of mild stroke patients

$$f'_k = r_k a_k, \qquad \forall k \neq i,$$

along with below constraints:

$$2f_k \leq z_k + (1 - a_k - t_k) \leq f_k + 1, \qquad \forall k \neq i,$$
$$2f'_k \leq r_k + a_k \leq f'_k + 1, \qquad \forall k \neq i.$$

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

$$\min \sum_{k \neq i} \pi_k x_k + \sum_{k \neq i} \lambda_k \left[\kappa_k t_k + \pi_k {W_k}^* f_k + \frac{\alpha}{\mu_k} f'_k \right]$$

Patient Type		Estimate	Std. Error	Z-Value	P-Value	95% CI
Severe Non-Stroke	**Month of the Year**					
	February	0.18	0.40	0.46	0.65	(-0.60,0.97)
	March	-0.69	0.50	-1.38	0.17	(-1.67,0.29)
	April	0.03	0.41	0.08	0.93	(-0.77,0.83)
	May	0.29	0.38	0.75	0.45	(-0.46,1.04)
	June	0.03	0.41	0.08	0.93	(-0.77,0.83)
	July	0.00	0.41	0.00	0.99	(-0.80,0.80)
	August	0.29	0.38	0.76	0.45	(-0.46,1.04)
	September	0.19	0.40	0.47	0.64	(-0.59,0.96)
	October	0.00	0.41	0.10	0.99	(-0.80,0.80)
	November	0.03	0.41	0.08	0.93	(-0.77,0.83)
	December	0.16	0.39	0.39	0.69	(-0.61,0.93)
	Day of the Week					
	Tuesday	0.13	0.30	0.44	0.66	(-0.45,0.72)
	Wednesday	-0.21	0.32	-0.64	0.52	(-0.85,0.43)
	Thursday	0.05	0.31	0.16	0.87	(-0.55,0.65)
	Friday	0.05	0.31	0.16	0.87	(-0.55,0.65)
	Saturday	0.10	0.30	0.33	0.74	(-0.50,0.69)
	Sunday	0.05	0.31	0.17	0.87	(-0.55,0.65)
	Hour of the Day					
	6 a.m.–12 p.m.	0.15	0.23	0.68	0.50	(-0.29,0.60)
	12 p.m.–6 p.m.	0.03	0.23	0.11	0.91	(-0.43,0.49)
	6 p.m.–12 a.m.	0.00	0.24	0.00	0.99	(-0.46,0.46)
	Constant	-3.51	0.38	-9.23	0.00	(-4.26,-2.77)

Table EC.3 Poisson regression analysis for estimating number of arrivals of severe non-stroke patients

$$
\begin{aligned}
&+ b_i\mu_i \left[(1-d_{ii})h_i(x_i, b_i-1) + d_{ii}h_i(x_i-1, b_i) + \sum_{j\neq i} d_{ij}\left(\frac{\alpha}{\mu_j}r_j - \pi_j {W_j}^* z'_j\right)\right] \\
&+ \sum_{k\neq i} b_k\mu_k \Bigg[(1-d_{kk})\left(-\frac{\alpha}{\mu_k}r'_k\right) + d_{ki}\Big(h_i(x_i-1, b_i+1) - h_i(x_i, b_i)\Big) \\
&- d_{kk}\left(\pi_k {W_k}^* z'_k\right) + \sum_{j\neq i,k} d_{kj}\left(\frac{\alpha}{\mu_j}r_j - \pi_j {W_j}^* z'_j\right)\Bigg]
\end{aligned}
$$

s.t.:

$$
\begin{aligned}
&\sum_{k\neq i} x_k \leq K - x_i, \quad \sum_{k\neq i} b_k \leq B - b_i, \\
&a_k \leq B - b_i, \quad x_i - K + 1 \leq a_k + t_k \leq 1, \qquad \forall k \neq i,
\end{aligned}
$$

Patient Type		Estimate	Std. Error	Z-Value	P-Value	95% CI
Severe Stroke	**Month of the Year**					
	February	0.30	0.45	0.66	0.51	(-0.58,1.18)
	March	0.20	0.45	0.43	0.67	(-0.69,1.08)
	April	-0.22	0.50	-0.44	0.66	(-1.21,0.77)
	May	0.10	0.46	0.21	0.83	(-0.80,1.00)
	June	0.14	0.46	0.30	0.77	(-0.76,1.04)
	July	0.36	0.43	0.83	0.41	(-0.49,1.21)
	August	-0.59	0.56	-1.06	0.29	(-1.68,0.50)
	September	0.32	0.44	0.72	0.47	(-0.55,1.18)
	October	0.20	0.45	0.43	0.66	(-0.69,1.08)
	November	0.23	0.45	0.51	0.61	(-0.65,1.11)
	December	0.37	0.43	0.85	0.40	(-0.48,1.22)
	Day of the Week					
	Tuesday	-0.49	0.38	-1.28	0.20	(-1.24,0.26)
	Wednesday	0.20	0.32	0.64	0.52	(-0.42,0.83)
	Thursday	-0.06	0.34	-0.18	0.86	(-0.72,0.60)
	Friday	-0.40	0.37	-1.07	0.28	(-1.13,0.33)
	Saturday	0.25	0.32	0.81	0.42	(-0.36,0.87)
	Sunday	0.11	0.32	0.34	0.74	(-0.53,0.75)
	Hour of the Day					
	6 a.m.–12 p.m.	-0.11	0.24	-0.47	0.64	(-0.58,0.36)
	12 p.m.–6 p.m.	-0.44	0.26	-1.66	0.10	(-0.95,0.08)
	6 p.m.–12 a.m.	-0.24	0.25	-0.98	0.33	(-0.73,0.24)
	Constant	-3.51	0.42	-8.36	0.00	(-4.34,-2.69)

Table EC.4 Poisson regression analysis for estimating number of arrivals of severe stroke patients

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

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

$$2f'_k \le r_k + a_k \le f'_k + 1, \qquad \forall k \ne i,$$

$$x_k \ge z_k L^*_k, \quad x_k \le (1 - z_k)(L^*_k - 1) + z_k M, \qquad \forall k \ne i,$$

$$x_k \ge z'_k(L^*_k + 1), \quad x_k \le (1 - z'_k)L^*_k + z'_k M, \qquad \forall k \ne i,$$

$$b_k \ge r_k \tilde{b}^*_k, \quad b_k \le (1 - r_k)(\tilde{b}^*_k - 1) + r_k M, \qquad \forall k \ne i,$$

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

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

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

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

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

$$
\begin{aligned}
&s_{ij} = d_{ij} r_j && 2s_{ij} \le d_{ij} + r_j \le s_{ij} + 1, && && \forall j \neq i,\\
&s'_{ij} = d_{ij} z'_j && 2s'_{ij} \le d_{ij} + z'_j \le s'_{ij} + 1, && && \forall j \neq i,\\
&e_k = d_{kk} r'_k && 2e_k \le d_{kk} + r'_k \le e_k + 1, && && \forall k \neq i,\\
&v_k = d_{kk} z'_k && 2v_k \le d_{kk} + z'_k \le v_k + 1, && && \forall k \neq i,\\
&u_{kj} = d_{kj} r_j && 2u_{kj} \le d_{kj} + r_j \le u_{kj} + 1, && \forall k \neq i, && j \neq i, k,\\
&y_{kj} = d_{kj} z'_j && 2y_{kj} \le d_{kj} + z'_j \le y_{kj} + 1, && \forall k \neq i, && j \neq i, k.
\end{aligned}
$$

The last set of variables is:

$$
\begin{aligned}
&m_k = b_k r'_k && m_k \le r'_k (B - b_i), m_k \le b_k, (r'_k - 1)M + b_k \le m_k, && && \forall k \neq i,\\
&m'_k = b_k e_k && m'_k \le e_k (B - b_i), m'_k \le b_k, (e_k - 1)M + b_k \le m'_k, && && \forall k \neq i,\\
&n_{ki} = b_k d_{ki} && n_{ki} \le d_{ki} (B - b_i), n_{ki} \le b_k, (d_{ki} - 1)M + b_k \le n_{ki}, && && \forall k \neq i,\\
&o_k = b_k v_k && o_k \le v_k (B - b_i), o_k \le b_k, (v_k - 1)M + b_k \le o_k, && && \forall k \neq i,\\
&p_{kj} = b_k u_{kj} && p_{kj} \le u_{kj} (B - b_i), p_{kj} \le b_k, (u_{kj} - 1)M + b_k \le p_{kj}, && \forall k \neq i, && j \neq i, k,\\
&q_{kj} = b_k y_{kj} && q_{kj} \le y_{kj} (B - b_i), q_{kj} \le b_k, , (q_{kj} - 1)M + b_k \le q_{kj}, && \forall k \neq i, && j \neq i, k.
\end{aligned}
$$

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

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

s.t.:

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

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

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

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

$$2f'_k \le r_k + a_k \le f'_k + 1, \qquad \forall k \ne i,$$

$$x_k \ge z_k L^*_k, \quad x_k \le (1 - z_k)(L^*_k - 1) + z_k M, \qquad \forall k \ne i,$$

$$x_k \ge z'_k(L^*_k + 1), \quad x_k \le (1 - z'_k)L^*_k + z'_k M, \qquad \forall k \ne i,$$

$$b_k \ge r_k \tilde{b}^*_k, \quad b_k \le (1 - r_k)(\tilde{b}^*_k - 1) + r_k M, \qquad \forall k \ne i,$$

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

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

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

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

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

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

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

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

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

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

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

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

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

$x_k, b_k, m_k, m'_k, o_k, n_{ki}, p_{kj}$ and q_{kj} are non-negative integer, $\forall k \ne i, \forall j \ne i, k,$

$a_k, t_k, z_k, z'_k, r_k, r'_k, f_k, f'_k, e_k, v_k, u_{kj}, y_{kj}$ are binary, $\forall k \ne i,$

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

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

EC.4. An example of non-linear cost functions

We provide an example of the convex piece-wise linear functions that were used in the analysis of Section 7.7. This example is dedicated to describing the waiting cost function for the mild stroke patients. However, the function structure for other patient types is similar. As we explained in that section the slope of the waiting cost function increases every three-hour interval and remains constant after the transfer threshold, which is 48 hours for the mild stroke patients. In Table EC.5,

we present the slopes in each time interval according to which the waiting cost is incurred. Note that the slope increases in equal increments up to one day (24 hours) of waiting. After that, the patient's health status presumably deteriorates even faster so that the increments in the slopes also increases at higher rates.

Wait Time Interval (Hours)	Waiting Cost per Hour	Wait Time Interval (Hours)	Waiting Cost per Hour
1–3	1.5	25–27	4.2
4–6	1.8	28–30	4.8
7–9	2.1	31–33	5.7
10–12	2.4	34–36	6.6
13–15	2.7	37–39	7.5
16–18	3.0	40–42	8.4
19–21	3.3	43–45	10.2
22–24	3.6	≥ 46	12

Table EC.5 An example of piece-wise linear function for patient waiting cost

For the purpose of comparison in Section 7.7 we have also taken another consideration into account when choosing these slopes. If we fit a linear function with zero intercept to the piecewise linear function described in Table EC.5 we will have the slope of the fitted line equal to 3.75 per hour. This is equivalent to waiting cost of 90 per day, which is the cost of waiting for the mild stroke patients in the case of linear costs. It is also noteworthy that when we assume non-linear functions for the cost of waiting we need to adjust the transfer cost accordingly. In this example, for the mild stroke patients the cost of waiting 48 hours according to the piece-wise linear function described in Table EC.5 is 239.40. This number should be considered as the transfer cost in this situation instead of 180 that was used when the costs were incurred linearly.

References

Porteus EL (2002) Foundations of Stochastic Inventory Theory (Stanford University Press).