Optimization
A JOURNAL OF MATHEMATICAL PROGRAMMING AND OPERATIONS RESEARCH

Taylor & Francis
Taylor & Francis Group

Optimization
A Journal of Mathematical Programming and Operations Research

ISSN: 0233-1934 (Print) 1029-4945 (Online) Journal homepage: www.tandfonline.com/journals/gopt20

Admission scheduling of inpatients by considering two inter-related resources: beds and operating rooms

Ting Zhu, Li Luo, Wenwu Shen, Xueru Xu & Ran Kou

To cite this article: Ting Zhu, Li Luo, Wenwu Shen, Xueru Xu & Ran Kou (2022) Admission scheduling of inpatients by considering two inter-related resources: beds and operating rooms, Optimization, 71:6, 1637-1675, DOI: 10.1080/02331934.2020.1829619

To link to this article: https://doi.org/10.1080/02331934.2020.1829619

Published online: 12 Oct 2020.

OPTIMIZATION
2022, VOL. 71, NO. 6, 1637–1675
https://doi.org/10.1080/02331934.2020.1829619

Taylor & Francis
Taylor & Francis Group

Admission scheduling of inpatients by considering two inter-related resources: beds and operating rooms

Ting Zhu[a], Li Luo[b], Wenwu Shen[c], Xueru Xu[b] and Ran Kou[b]

[a]West China Biomedical Big Data Center, West China Hospital/West China School of Medicine, Sichuan University, Chengdu, People's Republic of China; [b]Business School, Sichuan University, Chengdu, People's Republic of China; [c]West China Hospital, Sichuan University, Chengdu, People's Republic of China

ABSTRACT

This paper studies the admission scheduling problem with considering the capacity usage of two inter-related resources (beds and operating rooms) between three consecutive stages of care during surgical patients' admissions to Chinese public hospitals, namely (1) pre-surgical inpatient bed, (2) surgery, (3) post-surgical inpatient bed. Demand comes from two types of patients: (1) emergency patients, who arise randomly and have to be admitted immediately, and (2) elective patients, whose admissions can be scheduled. The authors develop a Markov Decision Process (MDP) model that decides how many elective patients should be admitted each day, with the objective of optimally using both operating room and inpatient bed capacity. The authors demonstrate that the number of elective admissions scheduled each day is monotonically increasing in the state of the system and in the bed capacity, indicating that a higher level of waiting elective patients and available (or total) bed capacity pulls more elective admissions through the system. The total discounted expected cost of the system exhibits decreasing marginal returns as the capacity in each stage increases independently of other stages. Through numerical experiments, there is substantial value in making admission scheduling decisions by jointly considering inpatient beds and operating rooms.

ARTICLE HISTORY
Received 8 January 2020
Accepted 18 September 2020

KEYWORDS
Operations management in health services; admission scheduling; capacity planning; operating room and inpatient bed; Markov decision process

2010 MATHEMATICS SUBJECT CLASSIFICATIONS
37Hxx Random dynamic systems; 37NxxApplications; 90Bxx Operations research and management science

1. Introduction

Patients' medical care usually involves multi-unit (upstream and downstream) in a single hospital, such as an emergency/outpatient care unit, a medical examination unit, an admission service unit, an operating room (OR), a central transportation unit, a postoperative anaesthesia care unit (PACU), an inpatient unit, an intensive care unit (ICU), a rehabilitation unit [1]. These units require a variety of medical resources (e.g. doctor, nurse, anaesthetist, technician, CT/MRI

CONTACT Wenwu Shen 573596113@qq.com

facility, bed, OR). Some surgeries like a brain tumour resection or cardiothoracic surgery procedures often imply a stay in the ICU post-surgery. It is now widely recognized that the integration of downstream resources such as beds or nursing care in ICUs leads to a better overall performance for surgeries [2]. More and more works address the admission scheduling problem of surgical patients considering the objective of jointly optimizing the utilization of several resources such as ORs, beds and nursing care [3]. By studying in isolation, the complexity and uncertainty that are inherent in healthcare systems may seem to be more manageable, but suboptimal conclusions may be drawn when the influences of other services are ignored or the impact of a change on the overall care chain is overlooked [4]. Failure to balance the use of capacity at consecutive stages to ensure smooth patient flow through the system often leads to blocking, cancellation, and inefficient use of capacity.

In North America and Europe, an elective surgical patient's admission date follows automatically from the date on which the patient's surgery is scheduled, namely elective surgical patients are always admitted on the day before or the day of their surgery (depending on the surgical procedure), they receive treatment first in ORs and then move to inpatient wards until being discharged. Usually, the doctor has reserved the OR time but cannot ensure that there are enough beds in the inpatient ward or downstream unit (PACU, ICU) to accommodate the post-operative patients [5]. The lack of beds in downstream units is the main reason for the cancellation of surgeries in the OR (62.5% of the total surgical cancellation [6]). Cancellation of the scheduled surgeries, therefore, results in inadequate utilization of OR time. In China, an elective surgical patient is first admitted to the hospital and then occupies an inpatient bed to finish all medical examinations that are needed for supporting the surgery before his or her surgery is scheduled. This process seems to be inefficient (with prolonged patient-in-bed-days prior to surgery) and differs considerably from the surgical inpatients admission process in North America and Europe, but it is unavoidable for the reason that the expense for all pre-surgical examinations can only be reimbursed after hospitalization. Since Chinese public hospitals are financially self-sufficient, the common scenario is that too many patients (both elective and emergency) are admitted to the hospital greedily for maximizing hospital revenue and minimizing patient waiting cost, causing tight or insufficient bed capacity in the ward. Without considering capacity shortage at the downstream OR stage, admitted preoperative patients are blocked in wards, occupying beds for multiple days. Making preoperative patients wait for an extended time decreases throughput rate of inpatient beds or generates overtime risk of ORs.

The allocation of hospital inpatient beds is a complex process, and directly affects a hospital's operational effectiveness and patient experience. In developed countries such as the United States, Singapore, the United Kingdom, and Canada, it shows that centralized and unified bed management by the Bed Management Unit (BMU) is the most effective way to optimize the allocation of beds [7].

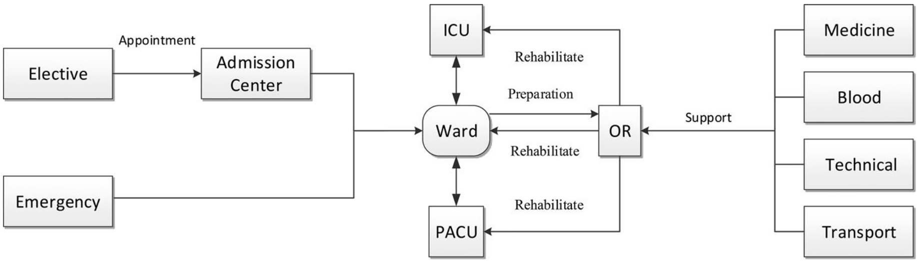


Figure 1. The conceptual framework of the hospitalization service units involved.

The centralized and unified allocation of beds needs to aggregate information about capacity usage at patients' multiple consecutive service stages, and comprehensively consider capacity interactive relationships between multiple units to achieve optimal operation of the whole system (see Figure 1). There has been a great imbalance between the hospital bed capacity and hospitalization demand in China, especially among large public hospitals. Hospital beds are critical resources but also limited because of the strict control of the expansion of public hospitals by the government health administration departments.

The impetus for this research problem comes from the recognition that beds, ORs, the necessary staffing and support for them are the main capacity constraints in public hospitals and the optimization of capacity planning requires one to proceed from the overall resource constraints of the hospital system, which ensures that patients receive timely, effective and high-quality inpatient medical services [8]. Our model is also motivated by the problems faced by a large public hospital in a major city in China, which we refer to as 'WCH' (the West China Hospital). The hospital operates a large inpatient department with a capacity of over 4300 licensed beds shared by 44 specialty care units. The limited supply of inpatient beds has led to overcrowding with 6000 elective patients waiting in the queue for admission every day. Waiting times for elective patients varied from several weeks to more than 1 year. Every day the bed manager of WCH faces the challenge of deciding the number of backlogged elective patients to be admitted, before accurate information about available capacity and demand is captured. The current practice in WCH is that bed dispatchers do not consider the OR capacity constraints when making decisions, and the OR scheduling does not consider the real-time occupancy of bed capacity in the ward. Faced with these problems, a more scientifically oriented allocation scheduling policy with considering the capacity usage of two inter-related resources (beds and operating rooms), which can be obtained by an analytical solution methodology, is greatly needed to serve as support in the decision-making process of patient admission scheduling.

In this paper, based on a dynamic multi-unit multi-stage admission control system of public hospitals in China, the admission scheduling of surgical patients considering the capacity allocation of two inter-related resources (inpatient bed and operating room) of the urology specialty care department (USCD) of WCH is studied, yet the proposed method can be applied to other settings. Patients progress through three consecutive service stages: admission-preoperative preparation stage (stage 0), intra-operative treatment stage (stage 1), and postoperative rehabilitation stage (stage 2). Each stage considers a resource constraint with finite capacity, particularly, the bottleneck resource consumed by patients in stage 0 are the preoperative beds, in stage 1 is the OR time, and in stage 2 is the postoperative beds. Two classes of patients arrive randomly on each day: elective and emergency. Emergency admission requests arise randomly and are assumed to enter the stage 0 and stage 1 on the day they become emergent. Elective patients need to book beds before admission and maybe backlogged on a waiting list. From this waiting list, a certain number of elective patients are chosen to be admitted each day, with each patient occupying a preoperative bed for a random duration of time. The elective patient admission decision comprehensively considers the balance and optimization of capacity utilization in three consecutive correlation units, especially involving the admission centre, ward and OR in Figure 1. On the day of admission, patients receive care first at stage 0 where they occupy beds in the ward for a random time, and then move to stage 1 by utilizing OR time for no more than one day, finally flow to stage 2 where they occupying ward beds for several days before discharge. Giving consideration to this downstream effect of the OR is essential for balancing the workload of the whole hospital. Clearly, the linked capacities of OR and bed impact the optimal number of elective patients that can be admitted on any given day. Under such settings, there are excessive patient waiting cost, OR and bed idling costs if the number of elective patients admitted to the hospital is too small; on the contrast, there are an over-scheduled cost of elective patients who are not satisfied on the admission day, OR overtime cost, and penalty cost for patients who cannot return to the ward after surgery. Our goal is to determine a dynamic admission policy for elective surgical patients to minimize the total discounted expected cost of the system over a finite horizon. The uniqueness of our problem is twofold: the perpetual uncertainty of the number of available beds when bed manager decides the number of elective patients to be admitted, and a system accounting for the linked capacity usage of three stages of service where inpatient bed provides care before OR.

Our model depicts a relatively simple service system with three sequential stages, two demand classes and two resource types. More importantly, our model can be used to approximate more complex multi-stage healthcare systems, especially when there are two obvious bottleneck stages and resources. Our contributions in this work can be summarized as follows. First, we formulate and analyse a dynamic admission scheduling model that integrates information about capacity usage at more than two service stages. Second, we prove that the number of elective patients scheduled to be admitted each day is monotone in the state variables and in the bed capacity, indicating that a higher level of waiting elective patients and available (or total) bed capacity pulls more elective patients through the system, thereby providing useful guidelines for adjusting scheduling decisions in practice. In addition, we show that the total discounted expected cost of the system exhibits decreasing marginal returns as the capacity in each stage increases independently of other stages. Finally, through numerical experiments based on data collected from WCH, we show that there is substantial value in making scheduling decisions considering two inter-related resources (bed and OR) compared to considering a single resource independently.

The rest of this paper is organized as follows. Section 2 reviews the relevant literature. Section 3 describes our model and formulation. Section 4 discusses the structural properties of the model and its optimal solutions. Section 5 transforms capacities at each stage as new decision variables, and studies their relationship to the optimal cost and the optimal scheduling decisions. Section 6 presents numerical results on the performance of joint scheduling. Section 7 discusses several extensions of our model and concludes our findings and potential implications.

2. Literature review

How to balance the use of capacity at different service stages to smooth patient flow is of concern to hospital administrators and researchers. The conflict is that the surgeons must schedule elective surgeries and book the operating theatre well in advance, and they usually assume that there will be an empty bed in the ward for the individual upon whom they will be operating. If not, the administrator is often forced to resolve the conflict through an unpleasant choice, and the path of least resistance is commonly to deny admission to an elective-surgery patient and, in effect, force the surgeon to cancel and reschedule the surgery. Such last-minute cancellations can have several negative consequences. In particular, they can wreak havoc with the schedules of the surgeons and the supporting staff, and require changes in the schedule of the operating theatre, since a cancelled surgery will eventually have to be rescheduled. Cancellations can also impose considerable stress on the patients and their families [9].

The complexity of the admission scheduling problem arises from the high degree of physical patient contact inherent in the delivery of health care services. Admission scheduling models attempt to maximize bed occupancy levels subject to three sources of variability: emergency admissions, patient length of stay, and patient service-mix requirements [10]. While hospitals desire high bed occupancy, slack capacity must be maintained for emergency admissions. A patient requires a bed and the use of other resources such as nursing care, surgical rooms, and support services; the demand for these resources is dependent on patient service mix. The models described above all focus on maximizing the utilization of bed resources, which can lead to extreme variations in the utilization of other resources. Ceschia and Schaerf [11] was the first to develop a hospital admission system based on nursing workload requirements. In his research, a simulation model was used to compare occupancy-related admission policies with systems that considered the relationship between admissions and nursing requirements. Ceschia and Schaerf found that workload variance was reduced when patients were scheduled by a workload-based model. In a similar study, Shukla [12] presented a scheduling system based on nursing workload and concluded that patient service mix had to be considered in the scheduling process in order to minimize changes in nursing staff assignment patterns. In contrast, a dynamic analytical model is formulated in this paper to solve the admission scheduling problem by balancing the use of capacity of ward and OR.

Moreover, the development of scheduling methods for multi-stage systems has remained a challenge and an open research area. Bowers [13] examined the balance between operating theatre and ICU bed capacity in a specialist facility providing elective heart and lung surgery. A simulation model was constructed to explore the interdependencies of resource availabilities and daily demand. This study provided an example of a capacity planning problem in which there is uncertainty in the demand and availabilities of two symbiotic resources. Liu et al. [1] introduced the first dynamic multi-day scheduling model that integrates information about capacity usage at two locations in a hospital. They considered a two-stage system in which patients receive surgeries at the upstream stage (OR) and may spend multiple days at the downstream stage (ward). However, our work features a three-stage service system in which OR has a downstream effect on the elective admission decision and simultaneously has an upstream effect on the ward bed occupancy. Kolker [14] developed a methodology aimed at answering a practical question: what maximum number of elective surgeries per day should be scheduled along with the competing demand from emergency surgeries in order to reduce diversion of an ICU with fixed bed capacity to an acceptable low level or prevent diversion at all. Min and Yih [15] proposed a stochastic scheduling model for elective surgeries that considers both uncertainty and downstream capacity constraints, for obtaining an optimal surgery schedule with respect to minimizing the sum of costs directly related to patients and expected overtime costs. The downstream capacities are modelled as constraints. In contrast to their approach, we focus on a dynamic admission scheduling policy in which OR capacity is considered as both upstream and downstream constraints.

A Master Surgical Schedule (MSS) specifies for each 'OR day' (i.e. operating room on a day) the planning cycle of recurring surgical procedure types that must be performed. van Oostrum et al. [16] considered the surgery planning problem that concerns the assignment of elective procedures to ORs on every day of the week. They next demonstrated that their approach is generic: it not only allows to level and control the workload of the involved surgical specialties, but also from succeeding departments such as ICUs and surgical wards. Vanberkel et al. [17] described an analytical approach to project the workload for downstream departments based on MSS too. Specifically, the ward occupancy distributions, patient admission/discharge distributions, and the distributions for ongoing interventions/treatments were computed. Recovering after surgery requires the support of multiple departments, such as nursing, physiotherapy, rehabilitation and long-term care. The model provides the foundation for a decision support tool to relate downstream hospital departments to the OR. Ma and Demeulemeester [18] used a multi-level integrative approach to determine the optimal patient mix and volume based on mathematical programming modelling and simulation analysis. The total expected bed shortage due to the variable length of stay of patients was minimized through reallocating the bed capacity and building balanced master surgery schedules. Fügener et al. [3] discussed the tactical MSS problem, concentrating on the effect that MSS has on the patient flow to downstream inpatient care units. They formulated a model to calculate the distributions of recovering patients in the downstream units expected from the MSS. Exact and heuristic algorithms for planning the MSS with the objective to minimize downstream costs were proposed by levelling bed demand and reducing weekend bed requests. van Essen et al. [19] developed an OR schedule by assigning OR blocks to a day in the planning horizon and incorporated both the randomness of the length of stay and of the bed occupancy to account for the variances in the downstream bed capacity. They proved the resulting problem to be NP-hard, but still modelled and solved the resulting problem as an Integer Linear Program (ILP), because their considered instances were small enough to be solved within a reasonable amount of time. Meanwhile, they also considered several what-if scenarios that relax some or all of the resource constraints. In previous works, both the static and dynamic multi-resource problem considered the surgical scheduling with bed, nursing staff, and other resources as a resource constraint, so patients receive care first in the OR. Our paper is distinct from these works in that patients receive care first in the ward then move to the OR service stage, so bed capacity is the major resource constraint with OR time as another constraint.

Samudra et al. [20] structurally classified the recent OR planning and scheduling literature into tables using patient type, used performance measures, decisions made, OR supporting units, uncertainty, research methodology and testing phase. Three levels of research were distinguished. The first level purely focuses on the OR department (including PACU and ICU). The second level targets the OR together with other areas that can be of interest in a hospital such as bed planning [21] or patient flow planning. The third level covers OR management in the broader context of patient care and therefore often includes different care services. In addition, the literature on primary surgical schedule and bed occupancy is divided into three categories: the impact of surgical scheduling on bed requirements [22], the establishment of models to predict bed occupancy rate but without considering surgical schedule [23], the use of bed capacity considered as a constraint on surgical scheduling rather than an optimization goal [24]. The existing research mainly considers the impact of surgical scheduling on the occupancy of downstream beds and the impact of demand for hospital beds on surgical scheduling, but less studies focus on joint scheduling and capacity planning of bed and OR. The shortcoming of current scheduling and capacity planning methods is that they formulate different units and resources in isolation [1]. However, in most medical systems, patients undergo a multi-stage treatment process and stay a random time in multiple medical units. If the downstream stage becomes fully occupied, access to it might be blocked for other patients upstream [25], thereby increasing the residence time in the unit and consuming more resources, and even causing the emergency patient to be unable to receive treatment in time. This congestion is in part attributable to surgical scheduling practices, which often focus on the efficient use of ORs but ignore resulting downstream bed utilization [26].

In a standard patient admission environment, new patients register at hospitals for treatment every day, meaning that patient admission scheduling problems are generally classified as dynamic. When patients require surgery and a room to stay for recovery, they must be assigned an admission day, surgery day and room. Due to the dynamic setting in which hospitals operate, it is extremely unlikely that a tactical level plan will be executed unchanged in practice. The problem addressed in this work concerns scheduling all patient admission requests once per day and continuously providing solutions day by day. Zhu et al. [27] studied the compatibility of short-term and long-term objectives in the context of the Dynamic Patient Admission Scheduling Problem (DPAS). A new short-term strategy that considers idle resource penalties and anticipatory information was presented for the problem. The resulting approach was then applied to the available DPAS benchmark, with its long-term solutions evaluated with respect to new lower bounds calculated using Dantzig–Wolfe decomposition and column generation. Ceschia and Schaerf [28] presented a further refinement of the PAS problem that includes, among other features, constraints on the utilization of operating rooms and a new model for managing patient delays. They performed an experimental analysis to tune the solver so as to identify the best configuration of the parameters with sufficient statistical confidence. In addition, they compared the results of the dynamic solver with those of the static one, for which all the events were known in advance. Lusby et al. [29] focused on providing a tool that can be used on a tactical planning level, where information about patients is gradually revealed over time. They also focused on the DPAS problem and devise an adaptive search (ALNS) procedure embedded within a simulated annealing (SA) framework to solve it. They tested and compared the proposed methodology on the large set of instances from [28].

3. Model

3.1. Study setting

Different types of demand have distinct admission appointment process, as shown in Figure 2. For hospitalization demand receiving primary services from the outpatient department (OD), they need to go to the admission service centre (ASC) for bed application. If at the moment a patient applies for a bed at ASC, there is an appropriate bed for him, he will be admitted directly and completes a series of admission procedures. The 'appropriate' means that the patient's acuity of illness is particularly high. If not, he will be placed on a waiting list until he is notified of the exact date of admittance by phone on the day before admission. For patients receiving primary services from the emergency department (ED), their admission decisions are made by both ED's physicians and the continuously treating physicians from a (or multiple) specialty care unit(s). If the hospitalization decision is determined after the multidisciplinary joint consultation, the bed application becomes valid. ASC's scheduler assigns beds to emergency patients as they become available.

Moreover, a description of the process for surgical admissions in China is shown in Figure 3, which is different from the surgical inpatient admission process in North America and Europe. In China, patients receive primary care from OD are called elective demand, and emergency demand for patients receive primary care from ED. All elective admission demand need bed application at ASC, where they register their individual information like demographic, disease diagnosis, source of payment, and so on. Bed allocation for elective demand and

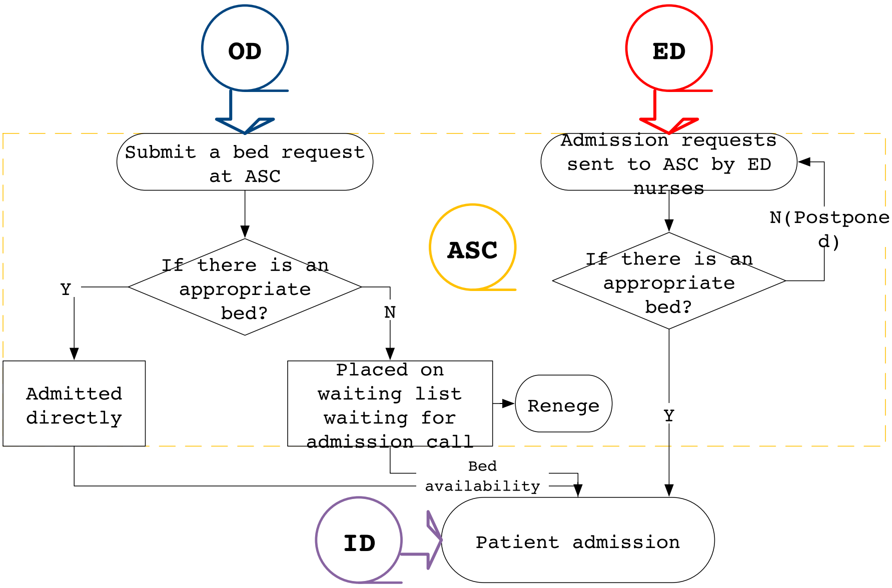

OD
ED
Submit a bed request at ASC
Admission requests sent to ASC by ED nurses
ASC
N(Postponed)
If there is an appropriate bed?
Y
N
If there is an appropriate bed?
Admitted directly
Placed on waiting list waiting for admission call
Renege
Y
Bed availability
ID
Patient admission

Figure 2. The admission appointment process of two types of patients.

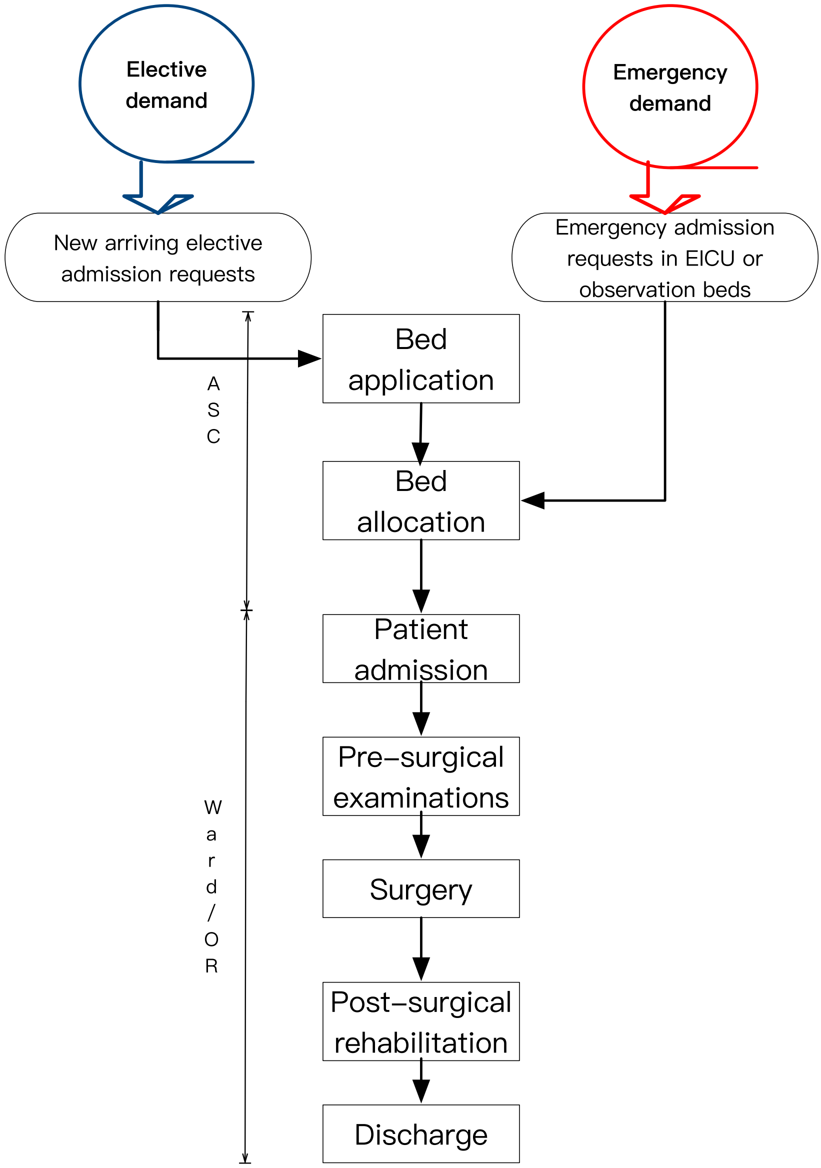

Elective demand
Emergency demand
New arriving elective admission requests
Emergency admission requests in EICU or observation beds
Bed application
Bed allocation
Patient admission
Pre–surgical examinations
Surgery
Post–surgical rehabilitation
Discharge
ASC
Ward/OR

Figure 3. The admission process of two types of patients.

emergency demand is conducted every day by ASC's bed planners. After the bed allocation process, patients are admitted to inpatient wards and go through the consecutive processes 'pre-surgical examinations', 'surgery' and 'post-surgical rehabilitation' until they are discharged.

3.2. *Problem description*

This section presents the model framework for our study. A dynamic admission control system for elective surgical patients is described. In order to correspond

Table 1. Summary of notations.

t	The decision periods. $t = 1, 2, 3, \dots, T$
$i \in \{0, 1, 2\}$	The service stages
w_t	State variable. The number of elective patients on the waiting list at the beginning of admission decision in period t
l_t	State variable. The number of unoccupied beds in ward at the beginning of period t
m_t	State variable. The number of elective patients undergoing surgery in period t
q_t	Decision variable. The number of scheduled elective patients to be admitted at the beginning of decision period t
$1 - \theta_t$	Random variable. The discharge rate of inpatients in period t. $\theta \in (0, 1)$
ξ_t	Random variable. The urgent rate of elective patients scheduled to be admitted at the beginning of period t
d_t	Random variable. The number of emergency patients who randomly arrive at hospital in period t
y_t	Random variable. The number of elective patients who randomly arrive at hospital in period t
γ	The per unit over-scheduled cost of elective patients
b	The per-unit waiting cost of elective patients
o	The per-unit overtime cost of operating room
c	The per-unit idle cost of operating room
p	The per-unit penalty cost for patients who are unable to return to ward after surgery
h	The per-unit holding cost of beds that have not been occupied
C^1	The total time capacity of an operating room
C^2	The total number of hospital beds of a ward
$n_t = C^2 - l_t$	The number of inpatients in the ward at the beginning of period t

to real-life, a scheduling and capacity planning model is formulated based on the operational practice of wards and operating rooms of WCH's USCD. The summary of notations can be seen in Table 1. All variables are non-negative integers and are related with the decision period t. All parameters $\gamma, b, o, c, p, h, C^1, C^2$ are positive integers and are independent over decision period t.

Consider a planning horizon of T days, numbered $t = 1, 2, 3, \dots, T$. Each decision epoch is defined as the time window from 16:00 to 18:00 of the decision day and that from 08:00 to 16:00 of the following admission day. Since the discharge-admission window is closed from 19:00 of the decision day to 07:00 of the admission day, this time window is not included in the decision epoch. The admission decision is determined in the afternoon of the previous day to provide sufficient time for patients' admission preparation. At the beginning of t (around 16:00 of each day), there are w_t elective patients backlogged on the waiting list (each patient waits for an available inpatient bed). Let l_t denote the total number of unoccupied beds in the care unit, n_t be the number of inpatients in the care unit at the beginning of t. m_t represents the number of elective surgeries planned to be performed in t and is decided by the OR scheduler of this care unit, the information on which is transferred to the bed scheduler of this care unit before an admission decision is made. Based on the observed information of w_t, l_t, n_t and m_t, elective patients to be admitted are first selected from the waiting list (denoted as q_t) and then are notified of admission. According to field observation and consultation with bed manager and corresponding planners, a percentage ξ_t of the q_t patients has to perform surgeries on the day they are admitted for their disease severity and urgency, we call them elective patients for emergency surgeries.

The remaining $(1-\xi_t)q_t$ patients wait in inpatient beds until their surgeries are scheduled.

Throughout the admission day, m_t planned elective surgeries are performed one by one according to their schedule; beds are released from discharged inpatients, which account for a fraction $(1-\theta_t)$ of (n_t-m_t). Demand for emergency and elective hospitalizations that arise over the day are non-negative integer-valued random variables denoted by d_t and y_t, respectively. We assume that d_t and y_t are independent and identically distributed (i.i.d.) over time, and bounded. Emergency demand must be admitted immediately and their surgeries must be performed on the same day when they arise, whereas elective demand may be waitlisted and their surgeries may be scheduled in the future. The scheduled q_t patients arrive at ASC. Each case for emergency surgery (both the $\xi_t q_t$ patients and the d_t patients) always proceeds through two main stages first in the hospital: the admission-preoperative preparation stage (stage 0), which takes place on the day the patient is admitted into the hospital; the intra-operative treatment stage (stage 1), in which the surgery is performed and the patient stays in stage 0 and stage 1 for no more than a day. Emergency surgeries can be inserted into the sequence of m_t undergoing elective surgeries planned in advance. On the other hand, each case of q_t for elective surgery needs to wait for available beds until 16:00 of the admission day. If bed capacity is insufficient for satisfying these $(1-\xi_t)q_t$ patients on the day, they are over-scheduled. These patients have to wait for next schedule days and beds either by living in hotels outside of the hospital or going back home directly, resulting in additional costs for patients, and their hospitalization requests are drawn back on the list. Otherwise, they receive service first in stage 0, where they finish the preoperative preparation and stay in ward for a random number of days before they are scheduled for surgery. After receiving surgery treatment service, patients will move to stage 2, called the post-operative rehabilitation stage, for recovery and observation. Patients stay in stage 2 for a random number of days before they are finally discharged.

We assume that each stage considers a resource constraint with finite capacity. For example, capacity in stage 1 might be measured by OR open time on any given day, denoted as C^1; let C^2 denote the total number of preoperative and postoperative beds in ward, respectively, measured in stage 0 and stage 2. Each patient consumes a random time of OR capacity in stage 1 and requires one bed in stage 0 or stage 2. We assume that each patient's surgery time is i.i.d. over the patient population and over time, and the total amount of OR capacity used by any number of surgical patients on any day can be approximately estimated by the convolution $S(\cdot)$. Within-day waiting time by patients or idling time by beds and surgeons that depend on the discharge-admission process and the sequencing of surgeries are not explicitly modelled, for negligible loss of goodwill in patients due to waiting and loss of revenue in the hospital due to idling. For the bottleneck of preoperative bed capacity in stage 0, each elective case that is still waitlisted incurs a waiting cost of b each day, and each scheduled elective case that is not satisfied on the admission day incurs an over-scheduled cost of γ each day. Intuitively, the per-unit cost for over-scheduling each patient is higher than that for making each patient wait a day, since each over-scheduled patient incurs not only a per-unit waiting cost but also additional transportation and accommodation costs. We assume $\gamma > b$. In stage 1, if more OR time is required than is available, then surge capacity is used incurring an OR overtime cost of o per unit; otherwise, an OR idling cost of c is incurred per unit. In stage 2, if more postoperative bed capacity is required than is available, postoperative patients have to occupy the post-anesthesia beds or unlicensed extra-beds for a while, incurring a penalty cost of p per unit; conversely, if more beds are available, a bed idling cost of h is generated per unit. Hospital revenue is generated by accepting patients, and daily unit bed cost has been defined as a performance metric to measure the opportunity revenue loss when keeping an inpatient bed idle for one day. Intuitively, we have $h > b$, to interpret the loss of idling a bed without fulfilling a waitlisted elective patient. After each decision epoch, both the waitlisted elective admission requests and the idled beds are transferred to the next epoch. The decision of q_t in stage 0 accounts for the usage of capacity in downstream stages and simultaneously impacts the overall system states, which finally cause the change of decisions on the number of waitlisted elective patients scheduled to be admitted in the next decision epochs.

The events in each epoch occur in the following sequence:

(1) At the beginning of epoch t, from 16:00 to 18:00 of the decision day, the bed dispatcher observes w_t elective patients backlogged on the waiting list, and (s)he is notified that there are l_t unoccupied beds (namely $n_t = C^2 - l_t$ inpatients) in the ward. Information about m_t patients to be performed surgeries on the following admission day is then captured from the surgical scheduling system. Based on above information, the bed dispatcher decides, out of the w_t elective patients, the number q_t to fulfill in t. Immediately after the decision, q_t patients are notified of admission one by one. Waiting costs are incurred for each of the $w_t - q_t$ waitlisted patients. We allow w_t, l_t, n_t, m_t and q_t to be non-negative real numbers.

(2) On the admission day, namely from 08:00 to 16:00 of t, m_t patients are moved to OR for surgery one by one according to their sequencing and timing of procedures. $(1 - \theta_t)(n_t - m_t)$ inpatients who are notified of discharge after doctor ward rounds release their beds. The q_t scheduled patients arrive at ASC. $\xi_t q_t$ urgent elective patients and a random number d_t of emergency patients are served at stage 0 immediately, their surgeries are randomly inserted into the m_t undergoing surgeries, bringing the total number of patients at stage 1 to $m_t + \xi_t q_t + d_t$. OR overtime and idling costs are incurred with $o[S(m_t + \xi_t q_t + d_t) - C^1]^+$ and $c[C^1 - S(m_t + \xi_t q_t + d_t)]^+$, respectively. $(1 - \xi_t)q_t$ patients wait for available beds until 16:00 of the admission day. Over-scheduled costs are incurred for

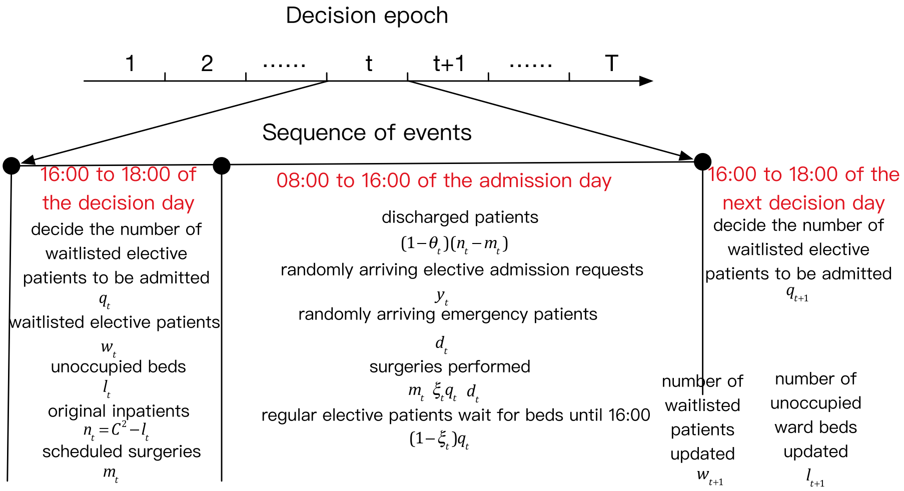

Decision epoch
1
2
……
t
t+1
……
T
Sequence of events
16:00 to 18:00 of the decision day
decide the number of waitlisted elective patients to be admitted
q_t
waitlisted elective patients
w_t
unoccupied beds
l_t
original inpatients
$n_t = C^2 - l_t$
scheduled surgeries
m_t
08:00 to 16:00 of the admission day
discharged patients
$(1-\theta_t)(n_t - m_t)$
randomly arriving elective admission requests
y_t
randomly arriving emergency patients
d_t
surgeries performed
m_t $\xi_t q_t$ d_t
regular elective patients wait for beds until 16:00
$(1-\xi_t)q_t$
16:00 to 18:00 of the next decision day
decide the number of waitlisted elective patients to be admitted
q_{t+1}
number of waitlisted patients updated
w_{t+1}
number of unoccupied ward beds updated
l_{t+1}

Figure 4. The sequence of events of the admission scheduling system.

each unfulfilled elective patients who are notified of admission today, namely, $\gamma[(1-\xi_t)q_t - (C^2 - \theta_t(n_t - m_t) - m_t - d_t - \xi_t q_t)^+]^+$. An additional random number y_t of new elective admission requests arises, bringing the total number of waitlisted patients to $w_t - q_t + [(1-\xi_t)q_t - (C^2 - \theta_t(n_t - m_t) - m_t - d_t - \xi_t q_t)^+]^+ + y_t$. We assume θ_t is i.i.d. over time, and the same for ξ_t.

(3) Each patient in stage 1 moves to stage 2. Penalty costs for postoperative patients who cannot move back to ward beds immediately are incurred with $p[\theta_t(n_t - m_t) + m_t + \xi_t q_t + d_t - C^2]^+$. Conversely, bed idling costs are incurred for each of the $[C^2 - \theta_t(n_t - m_t) - m_t - q_t - d_t]^+$ unoccupied beds.

(4) After 16:00 of the admission day, the system proceeds into the next decision epoch.

The objective of the problem is to determine a dynamic scheduling policy that minimizes the total discounted expected cost of the system over the planning horizon. The decision process is illustrated in Figure 4.

3.3. *Dynamic programming formulation*

The problem introduced above can be formulated as a Markov Decision Process (MDP). If too few elective patients are scheduled to be admitted in a day, it risks incurring high waiting costs for waitlisted elective patients and high idling costs for both OR time and inpatient bed. In contrast, if too many elective patients are scheduled, patient waiting costs and resource idling costs are reduced but over-scheduled costs for scheduled elective patients and surge capacity costs for both

OR time and bed might be high. Therefore, the admission scheduling decision needs to optimize and balance the use of capacity system-wide.

Recall that the decision to make in each epoch t is the number of waitlisted elective patients q_t to admit. The state of the system before the decision q_t is made is denoted as (w_t, m_t, l_t), indicating the state of the three stages (stage 0, stage 1 and stage 2), respectively. The decision q_t is constrained by $0 \leq q_t \leq w_t$, because the number to be scheduled cannot exceed the number currently on the waiting list. Let $x_t = (C^2 - \theta_t(C^2 - l_t - m_t) - m_t - d_t - \xi_t q_t)^+$ represent the number of beds remained after all primary inpatients and postoperative patients are served by beds, where $C^2 - l_t = n_t$.

The system evolves as follows:

$$\begin{aligned} w_{t+1} &= w_t - q_t + y_t + [(1-\xi_t)q_t - (C^2 - \theta_t(C^2 - l_t - m_t) \\ &\quad - m_t - d_t - \xi_t q_t)^+]^+ \\ &= w_t - q_t + y_t + [(1-\xi_t)q_t - x_t]^+ \qquad (1) \\ l_{t+1} &= C^2 - \min\{C^2, \theta_t(C^2 - l_t - m_t) + m_t + d_t + \xi_t q_t\} \\ &\quad - \min\{(1-\xi_t)q_t, ((1-\theta_t)C^2 \\ &\quad + \theta_t l_t - (1-\theta_t)m_t - d_t - \xi_t q_t)^+\} \\ &= x_t - (1-\xi_t)q_t + [(1-\xi_t)q_t - x_t]^+ \qquad (2) \end{aligned}$$

The single-epoch cost function can be written as

$$\begin{aligned} \hat{C}_t(q_t, w_t, l_t, m_t) &= b(w_t - q_t) + \gamma[(1-\xi_t)q_t - x_t]^+ \\ &\quad + o[S(m_t + \xi_t q_t + d_t) - C^1]^+ \\ &\quad + c[C^1 - S(m_t + \xi_t q_t + d_t)]^+ \\ &\quad + p[\theta_t(C^2 - l_t - m_t) + m_t + \xi_t q_t + d_t - C^2]^+ \\ &\quad + h[C^2 - \theta_t(C^2 - l_t - m_t) - m_t - q_t - d_t]^+ \qquad (3) \\ C_t(q_t, w_t, l_t, m_t) &= \mathbb{E}[\hat{C}_t(q_t, w_t, l_t, m_t)] \qquad (4) \end{aligned}$$

where $(\cdot)^+ = \max(\cdot, 0)$, $\mathbb{E}[\cdot]$ is to obtain the expectation. In Equation (3), the first term captures the waiting costs for backlogged elective patients in epoch t; the second term evaluates the over-scheduled costs for elective patients who are scheduled of admission but do not be served for shortage of bed capacity; the third and fourth terms characterize the overtime and idling costs for OR time, respectively; the last two terms compute the penalty costs for postoperative patients who cannot move back to ward beds immediately and bed idling costs, respectively.

Let $U_t(q_t, w_t, l_t, m_t)$ denote the total discounted cost incurred from epochs t to T when the state just before the decision is made in t is captured by (w_t, l_t, m_t). Let $V_t(w_t, l_t, m_t)$ be the optimal value of $U_t(q_t, w_t, l_t, m_t)$, assuming a discount factor of $\alpha \in (0, 1)$. The Bellman equation can be written as follows:

$$U_t(q_t, w_t, l_t, m_t) = C_t(q_t, w_t, l_t, m_t) + \alpha\mathbb{E}[V_{t+1}(w_{t+1}, l_{t+1}, m_{t+1})|(w_t, l_t, m_t)] \tag{5}$$

$$V_t(w_t, l_t, m_t) = \min_{0\leq q_t\leq w_t} U_t(q_t, w_t, l_t, m_t) \tag{6}$$

We take the termination function when $T < \infty$ to be $V_{T+1}(\cdot, \cdot, \cdot) = 0$.

Recall that (w_t, m_t, l_t) tracks the number of waitlisted elective patients at stage 0, the number of scheduled elective surgeries at stage 1 and the number of unoccupied beds at stage 2, analysing the relationship between the three state variables and the decision variable q_t is an effective way to optimize the decisions of q_t. As the number of elective patients on the waiting list w_t increases, intuition seems to suggest that more patients should be admitted, i.e. q_t should be larger, to avoid excessive waiting at stage 0. However, q_t might increase or decrease when w_t increases, depending on the usage of capacity at both downstream stages. As the number of unoccupied beds l_t increases, q_t might be larger to reduce idling cost of beds at stage 2. However, more patients pulled from the waiting list through the system might increase the overtime risk of OR. As the number of scheduled elective surgeries m_t increases, it intuitively suggests that fewer patients should be admitted, i.e. q_t should be smaller, to reduce overtime cost of OR at stage 1, however, it might result in excessive waiting of patients at stage 0 and idling cost of beds at stage 2, respectively. The adjustment of decisions q_t with changes of m_t depends on the relative value of m_t compared to the value of capacities C^1 and C^2.

When the number m_t of scheduled elective surgeries is small compared to the OR capacity C^1, it is crucial to reduce the idling cost of OR time. Hence, more patients are informed of admission even if it leads to shortage risk of inpatient beds at stage 2. As m_t increases, reducing the shortage risk of inpatient beds becomes more important, fewer patients should be informed of admission even if idling cost might incur at stage 1 and excessive waiting of patients might occur at stage 0. However, when m_t is sufficiently large, the overtime cost at stage 1 is almost the same for any newly admitted patient. In this case, balancing costs incurred at stage 0 and at stage 2 is of more significance. When m_t is very small compared to the bed capacity C^2, q_t should be larger to reduce waiting cost of patients at stage 0 and idling cost of beds at stage 2; on the contrary, q_t should be fewer. The above examples show that it is important to explore the relationships among the model variables to provide a clear direction on how to adjust decisions as the system state changes.

4. Structural properties of optimal solutions

4.1. Structural properties of the single-epoch problem

By analysing the relationship between the decision variable q_t and the system state (w_t, m_t, l_t) of the single-epoch cost function (4), we derive the structural properties that will elucidate the characteristics of the optimal policies to solve the single-epoch problem. We first prove the convexity of the single-epoch cost function and state transition functions.

Lemma 4.1: *For any d_t, y_t, θ_t, ξ_t, the following properties of the single-epoch cost function $C_t(q_t, w_t, l_t, m_t)$ (4) hold:*

(1) *For any given (l_t, m_t), $C_t(q_t, w_t, l_t, m_t)$ is jointly convex in (q_t, w_t).*
(2) *For any given (w_t, m_t), $C_t(q_t, w_t, l_t, m_t)$ is jointly convex in (q_t, l_t).*
(3) *For any given (w_t, l_t), $C_t(q_t, w_t, l_t, m_t)$ is jointly convex in (q_t, m_t).*
(4) *For any given (q_t, l_t), $C_t(q_t, w_t, l_t, m_t)$ is jointly convex in (w_t, m_t).*
(5) *For any given (q_t, m_t), $C_t(q_t, w_t, l_t, m_t)$ is jointly convex in (w_t, l_t).*
(6) *For any given (q_t, w_t), $C_t(q_t, w_t, l_t, m_t)$ is jointly convex in (m_t, l_t).*

Proof: (a) The convexity of $C_t(q_t, w_t, l_t, m_t)$ in (q_t, w_t) is discussed in two cases.

First, when $x_t = (C^2 - \theta_t(C^2 - l_t - m_t) - m_t - d_t - \xi_t q_t)^+ > 0$, we have

$$\begin{aligned}\hat{C}_t(q_t, w_t, l_t, m_t) = {} & b(w_t - q_t) + \gamma[(1-\theta_t)m_t + d_t + q_t - (1-\theta_t)C^2 - \theta_t l_t]^+ \\ & + o[S(m_t + \xi_t q_t + d_t) - C^1]^+ + c[C^1 - S(m_t + \xi_t q_t + d_t)]^+ \\ & + h[C^2 - \theta_t(C^2 - l_t - m_t) - m_t - q_t - d_t]^+ \end{aligned} \tag{7}$$

The first term $b(w_t - q_t)$ is convex in q_t. Then we discuss the two cases when $(1 - \theta_t)C^2 + \theta_t l_t - (1 - \theta_t)m_t - d_t - q_t \le 0$ and $(1 - \theta_t)C^2 + \theta_t l_t - (1 - \theta_t)m_t - d_t - q_t > 0$.

(i) When $(1 - \theta_t)C^2 + \theta_t l_t - (1 - \theta_t)m_t - d_t - q_t \le 0$ holds, we have

$$\begin{aligned}& \hat{C}_t(q_t, w_t, l_t, m_t) \\ & \quad = b(w_t - q_t) + \gamma[(1-\theta_t)m_t + d_t + q_t - (1-\theta_t)C^2 - \theta_t l_t] \\ & \qquad + o[S(m_t + \xi_t q_t + d_t) - C^1]^+ + c[C^1 - S(m_t + \xi_t q_t + d_t)]^+ \end{aligned} \tag{8}$$

(ii) When $(1 - \theta_t)C^2 + \theta_t l_t - (1 - \theta_t)m_t - d_t - q_t > 0$ holds, we have

$$\begin{aligned}\hat{C}_t(q_t, w_t, l_t, m_t) = {} & b(w_t - q_t) + o[S(m_t + \xi_t q_t + d_t) - C^1]^+ \\ & + c[C^1 - S(m_t + \xi_t q_t + d_t)]^+ \\ & + h[(1-\theta_t)C^2 + \theta_t l_t - (1-\theta_t)m_t - d_t - q_t] \end{aligned} \tag{9}$$

Since $S(k)$ represents the sum of k independently identical distributed (i.i.d.) non-negative random variables, i.e. the sum of k surgeries' operation time, $\{S(k), k = 0, 1, 2, \ldots\}$ is stochastic increasing and linear in sample path sense. Thus, if g is a convex function, $g(S(k))$ is convex in k. As $o[S(m_t + \xi_t q_t + d_t) - C^1]^+ + c[C^1 - S(m_t + \xi_t q_t + d_t)]^+$ is convex in $S(m_t + \xi_t q_t + d_t)$, and $S(m_t + \xi_t q_t + d_t)$ is stochastic increasing and linear in q_t, the convexity of $o[S(m_t + \xi_t q_t + d_t) - C^1]^+ + c[C^1 - S(m_t + \xi_t q_t + d_t)]^+$ in q_t holds. Moreover, due to the fact that the convexity of the sum of multiple convex functions is preserved, the convexity of $\hat{C}_t(q_t, w_t, l_t, m_t)$ in q_t holds. In Equation (7), only the first term includes w_t, the convexity of $\hat{C}_t(q_t, w_t, l_t, m_t)$ in w_t holds intuitively. In addition, due to the fact that the convexity of a real function is preserved under linear transformation of its arguments (see Theorem 5.7 in [30]), we know that $\hat{C}_t(q_t, w_t, l_t, m_t)$ is jointly convex in (q_t, w_t).

Second, when $x_t = (C^2 - \theta_t(C^2 - l_t - m_t) - m_t - d_t - \xi_t q_t)^+ = 0$, i.e. $C^2 - \theta_t(C^2 - l_t - m_t) - m_t - d_t - \xi_t q_t \leq 0$, we have

$$\begin{aligned}\hat{C}_t(q_t, w_t, l_t, m_t) = {} & b(w_t - q_t) + \gamma(1 - \xi_t)q_t \\ & + o[S(m_t + \xi_t q_t + d_t) - C^1]^+ + c[C^1 - S(m_t + \xi_t q_t + d_t)]^+ \\ & + p[\theta_t(C^2 - l_t - m_t) + m_t + \xi_t q_t + d_t - C^2] \end{aligned} \tag{10}$$

According to the above mentioned, $o[S(m_t + \xi_t q_t + d_t) - C^1]^+ + c[C^1 - S(m_t + \xi_t q_t + d_t)]^+$ is convex in q_t. In Equation (10), the first, second and fifth terms are linear in q_t, respectively. Hence, $\hat{C}_t(q_t, w_t, l_t, m_t)$ is convex in q_t. Intuitively, $\hat{C}_t(q_t, w_t, l_t, m_t)$ is jointly convex in (q_t, w_t).

In summary, $C_t(q_t, w_t, l_t, m_t)$ is jointly convex in (q_t, w_t).

(b) The convexity of $C_t(q_t, w_t, l_t, m_t)$ in (q_t, l_t) is discussed in two cases.

First, when $x_t = (C^2 - \theta_t(C^2 - l_t - m_t) - m_t - d_t - \xi_t q_t)^+ > 0$, we have proved the convexity of $\hat{C}_t(q_t, w_t, l_t, m_t)$ in q_t in (a). According to Equation (7), the second and fifth terms are related to l_t. When $(1 - \theta_t)C^2 + \theta_t l_t - (1 - \theta_t)m_t - d_t - q_t \leq 0$ holds, i.e. $l_t \in \frac{1}{\theta_t}[(1 - \theta_t)m_t + d_t + \xi_t q_t - (1 - \theta_t)C^2, (1 - \theta_t)m_t + d_t + q_t - (1 - \theta_t)C^2]$, $\hat{C}_t(q_t, w_t, l_t, m_t)$ is decreasing convex in l_t. When $(1 - \theta_t)C^2 + \theta_t l_t - (1 - \theta_t)m_t - d_t - q_t > 0$ holds, i.e. $l_t \in \frac{1}{\theta_t}[(1 - \theta_t)m_t + d_t + q_t - (1 - \theta_t)C^2, C^2]$, $\hat{C}_t(q_t, w_t, l_t, m_t)$ is increasing convex in l_t. Hence, $\hat{C}_t(q_t, w_t, l_t, m_t)$ is convex in l_t. Similarly, $\hat{C}_t(q_t, w_t, l_t, m_t)$ is jointly convex in (q_t, l_t) (see Theorem 5.7 in [30]).

Second, when $x_t = (C^2 - \theta_t(C^2 - l_t - m_t) - m_t - d_t - \xi_t q_t)^+ = 0$, i.e. $C^2 - \theta_t(C^2 - l_t - m_t) - m_t - d_t - \xi_t q_t \leq 0$, we have proved the convexity of $\hat{C}_t(q_t, w_t, l_t, m_t)$ in q_t in (a). According to Equation (10), only the fifth term is related to l_t. Clearly, $\hat{C}_t(q_t, w_t, l_t, m_t)$ is convex in l_t. Hence, $\hat{C}_t(q_t, w_t, l_t, m_t)$ is jointly convex in (q_t, l_t).

In summary, $C_t(q_t, w_t, l_t, m_t)$ is jointly convex in (q_t, l_t).

(c) The convexity of $C_t(q_t, w_t, l_t, m_t)$ in (q_t, m_t) is discussed in two cases.

First, when $x_t = (C^2 - \theta_t(C^2 - l_t - m_t) - m_t - d_t - \xi_t q_t)^+ > 0$, we have proved the convexity of $\hat{C}_t(q_t, w_t, l_t, m_t)$ in q_t in (a). According to Equation (7), all terms except for the first term are related to m_t. When $(1-\theta_t)C^2 + \theta_t l_t - (1-\theta_t)m_t - d_t - q_t \le 0$ holds, i.e. $m_t \in \frac{1}{1-\theta_t}[(1-\theta_t)C^2 + \theta_t l_t - d_t - q_t, (1-\theta_t)C^2 + \theta_t l_t - d_t - \xi_t q_t]$, $\hat{C}_t(q_t, w_t, l_t, m_t)$ is increasing convex in m_t. Conversely, when $(1-\theta_t)C^2 + \theta_t l_t - (1-\theta_t)m_t - d_t - q_t > 0$ holds, i.e. $m_t \in \frac{1}{1-\theta_t}[0, (1-\theta_t)C^2 + \theta_t l_t - d_t - q_t]$, $\hat{C}_t(q_t, w_t, l_t, m_t)$ is decreasing convex in m_t. $\hat{C}_t(q_t, w_t, l_t, m_t)$ is jointly convex in (q_t, m_t) (see Theorem 5.7 in [30]).

Second, when $x_t = (C^2 - \theta_t(C^2 - l_t - m_t) - m_t - d_t - \xi_t q_t)^+ = 0$, i.e. $C^2 - \theta_t(C^2 - l_t - m_t) - m_t - d_t - \xi_t q_t \le 0$, we have proved the convexity of $\hat{C}_t(q_t, w_t, l_t, m_t)$ in q_t in (a). According to Equation (10), the third, fourth and fifth terms are related to m_t. Clearly, $\hat{C}_t(q_t, w_t, l_t, m_t)$ is convex in m_t. Hence, $\hat{C}_t(q_t, w_t, l_t, m_t)$ is jointly convex in (q_t, m_t).

In summary, $C_t(q_t, w_t, l_t, m_t)$ is jointly convex in (q_t, m_t).

Similar to the proof of (a), (b) and (c), it is easy to prove the convexity of $C_t(q_t, w_t, l_t, m_t)$ in (w_t, m_t), (w_t, l_t), and (m_t, l_t). ■

Lemma 4.2: *For any d_t, y_t, θ_t, ξ_t, w_{t+1} and l_{t+1} are jointly convex in their arguments (q_t, w_t, l_t, m_t), respectively.*

Proof: (a) According to Equation (1), i.e. $w_{t+1} = w_t - q_t + y_t + [(1-\xi_t)q_t - x_t]^+$, First, when $x_t = (C^2 - \theta_t(C^2 - l_t - m_t) - m_t - d_t - \xi_t q_t)^+ > 0$, we have $w_{t+1} = w_t - q_t + y_t + [(1-\theta_t)m_t + d_t + q_t - (1-\theta_t)C^2 - \theta_t l_t]^+$. If $(1-\theta_t)m_t + d_t + q_t - (1-\theta_t)C^2 - \theta_t l_t > 0$, we have $w_{t+1} = w_t + y_t + (1-\theta_t)m_t + d_t - (1-\theta_t)C^2 - \theta_t l_t$; conversely, $w_{t+1} = w_t - q_t + y_t$. Hence, w_{t+1} is convex in q_t, w_t, m_t and l_t, respectively.

Second, when $x_t = (C^2 - \theta_t(C^2 - l_t - m_t) - m_t - d_t - \xi_t q_t)^+ = 0$, $w_{t+1} = w_t - \xi_t q_t + y_t$ is convex in q_t, w_t, m_t and l_t, respectively.

In summary, w_{t+1} is jointly convex in their arguments (q_t, w_t, l_t, m_t).

(b) According to Equation (2), i.e. $l_{t+1} = x_t - (1-\xi_t)q_t + [(1-\xi_t)q_t - x_t]^+$.

First, when $x_t = (C^2 - \theta_t(C^2 - l_t - m_t) - m_t - d_t - \xi_t q_t)^+ > 0$, we have $l_{t+1} = (1-\theta_t)C^2 + \theta_t l_t - (1-\theta_t)m_t - d_t - q_t + ((1-\theta_t)m_t + d_t + q_t - (1-\theta_t)C^2 - \theta_t l_t)^+$. If $(1-\theta_t)m_t + d_t + q_t - (1-\theta_t)C^2 - \theta_t l_t > 0$, we have $l_{t+1} = 0$; conversely, $l_{t+1} = (1-\theta_t)C^2 + \theta_t l_t - (1-\theta_t)m_t - d_t - q_t$. Hence, l_{t+1} is convex in q_t, w_t, m_t and l_t, respectively.

Second, when $x_t = (C^2 - \theta_t(C^2 - l_t - m_t) - m_t - d_t - \xi_t q_t)^+ = 0$, $l_{t+1} = 0$.

In summary, l_{t+1} is jointly convex in their arguments (q_t, w_t, l_t, m_t). ■

4.2. Structural properties of the multi-epoch problem

According to the properties proved in Section 4.1, we derive the structural properties that will elucidate the characteristics of the optimal policies to solve the dynamic multi-epoch problem, thus providing decision-makers with helpful guidance on these policies.

Lemma 4.3: *For any d_t, y_t, θ_t, ξ_t, the following properties of the multi-epoch cost function $U_t(q_t, w_t, l_t, m_t)$ (5) and $V_t(w_t, l_t, m_t)$ (6) hold:*

(1) *For any given (l_t, m_t), $U_t(q_t, w_t, l_t, m_t)$ is jointly convex in (q_t, w_t).*
(2) *For any given (w_t, m_t), $U_t(q_t, w_t, l_t, m_t)$ is jointly convex in (q_t, l_t).*
(3) *For any given (w_t, l_t), $U_t(q_t, w_t, l_t, m_t)$ is jointly convex in (q_t, m_t).*
(4) *For any given (q_t, l_t), $U_t(q_t, w_t, l_t, m_t)$ is jointly convex in (w_t, m_t).*
(5) *For any given (q_t, m_t), $U_t(q_t, w_t, l_t, m_t)$ is jointly convex in (w_t, l_t).*
(6) *For any given (q_t, w_t), $U_t(q_t, w_t, l_t, m_t)$ is jointly convex in (m_t, l_t).*
(7) *For any given m_t, $V_t(w_t, l_t, m_t)$ is jointly convex in (w_t, l_t).*
(8) *For any given l_t, $V_t(w_t, l_t, m_t)$ is jointly convex in (w_t, m_t).*
(9) *For any given w_t, $V_t(w_t, l_t, m_t)$ is jointly convex in (m_t, l_t).*

Proof: We prove the above properties using methods of mathematical induction. Assuming that $V_{t+1}(w_{t+1}, l_{t+1}, m_{t+1})$ is jointly convex in its arguments (which is true for $t = T$). We know that w_{t+1} and l_{t+1} are jointly convex in their arguments (q_t, w_t, l_t, m_t), respectively. Hence, $\mathbb{E}[V_{t+1}(w_{t+1}, l_{t+1}, m_{t+1})]$ is jointly convex in its arguments. Therefore, $U_t(q_t, w_t, l_t, m_t)$ is jointly convex in its arguments. According to Equation (6), $V_t(w_t, l_t, m_t)$ is obtained by minimizing a convex function $U_t(q_t, w_t, l_t, m_t)$ in a convex feasible region, it follows that $V_t(w_t, l_t, m_t)$ is also convex (see Theorem A.4 in [31]). By induction, the property holds for all t. ■

Theorem 4.4: *For any d_t, y_t, θ_t, ξ_t, the following properties of $C_t(q_t, w_t, l_t, m_t)$ (4), $U_t(q_t, w_t, l_t, m_t)$ (5) and $V_t(w_t, l_t, m_t)$ (6) hold:*

(1) *$C_t(q_t, w_t, l_t, m_t)$ is submodular in (q_t, w_t, l_t).*
(2) *$C_t(q_t, w_t, l_t, m_t)$ is submodular in (m_t, w_t, l_t).*
(3) *$U_t(q_t, w_t, l_t, m_t)$ is submodular in (q_t, w_t, l_t).*
(4) *$V_t(w_t, l_t, m_t)$ is submodular in (w_t, l_t).*

Proof: (a) See LEMMA EC.1 of the Online Appendix in Huh et al. [31], we know that if $V(z)$ is a convex function in z and a is a positive constant, then $V(x - ay)$ is submodular in x and y. For any given m_t, when $x_t = (C^2 - \theta_t(C^2 - l_t - m_t) - m_t - d_t - \xi_t q_t)^+ > 0$, for Equation (7), we know its first term $b(w_t - q_t)$ is linear and is submodular in q_t, w_t and l_t; its second and fifth terms are jointly convex in l_t and q_t, and thus are submodular in q_t, w_t and l_t, respectively; its third and fourth terms are convex in q_t and are submodular in q_t, w_t and l_t, respectively. Otherwise, when $x_t = 0$, for Equation (10), clearly its first term is submodular in q_t, w_t and l_t; its second, third and fourth terms are convex in q_t and are submodular in q_t, w_t and l_t, respectively; its fifth term is linear in l_t and q_t, and thus is submodular in q_t, w_t and l_t, which complete the proof of (1). Similarly, the property of (2) is easy to prove.

(c) Since the properties of (1) and (2) hold, to prove (3), we only need to prove that for any d_t, y_t, θ_t, ξ_t, $V_{t+1}(w_{t+1}, l_{t+1}, m_{t+1})$ is submodular in w_t and l_t. Assuming that $V_{t+1}(w_{t+1}, l_{t+1}, m_{t+1})$ is submodular in w_{t+1} and l_{t+1}. For easy writing, we remove m_{t+1} from $V_{t+1}(w_{t+1}, l_{t+1}, m_{t+1})$. According to Equations (1) and (2), we have $V_{t+1}(w_{t+1}, l_{t+1}) = V_{t+1}(w_t - q_t + y_t + [(1-\xi_t)q_t - x_t]^+, x_t - (1-\xi_t)q_t + [(1-\xi_t)q_t - x_t]^+)$, where $x_t = (C^2 - \theta_t(C^2 - l_t - m_t) - m_t - d_t - \xi_t q_t)^+$. We proof the submodularity of $V_{t+1}(w_{t+1}, l_{t+1}, m_{t+1})$ in w_{t+1} and l_{t+1} in two cases.

(i) When $x_t = C^2 - \theta_t(C^2 - l_t - m_t) - m_t - d_t - \xi_t q_t > 0$ holds, we have $w_{t+1} = w_t - q_t + y_t + [(1-\theta_t)m_t + d_t + q_t - (1-\theta_t)C^2 - \theta_t l_t]^+$ and $l_{t+1} = (1-\theta_t)C^2 + \theta_t l_t - (1-\theta_t)m_t - d_t - q_t + ((1-\theta_t)m_t + d_t + q_t - (1-\theta_t)C^2 - \theta_t l_t)^+$. (a) If $(1-\theta_t)m_t + d_t + q_t - (1-\theta_t)C^2 - \theta_t l_t > 0$, we have $w_{t+1} = w_t + y_t + (1-\theta_t)m_t + d_t - (1-\theta_t)C^2 - \theta_t l_t$ and $l_{t+1} = 0$; (b) conversely, $w_{t+1} = w_t - q_t + y_t$ and $l_{t+1} = (1-\theta_t)C^2 + \theta_t l_t - (1-\theta_t)m_t - d_t - q_t$.

(ii) When $x_t = (C^2 - \theta_t(C^2 - l_t - m_t) - m_t - d_t - \xi_t q_t)^+ = 0$ holds, $w_{t+1} = w_t - \xi_t q_t + y_t$ and $l_{t+1} = 0$.

For (i)(a) with any $w_t^+ \geq w_t^-$ and $l_t^+ \geq l_t^-$, we have

$$\begin{aligned} &V_{t+1}(w_t^+ + y_t + (1-\theta_t)m_t + d_t - (1-\theta_t)C^2 - \theta_t l_t^+, 0)\\ &\quad - V_{t+1}(w_t^+ + y_t + (1-\theta_t)m_t + d_t - (1-\theta_t)C^2 - \theta_t l_t^-, 0)\\ &\quad - V_{t+1}(w_t^- + y_t + (1-\theta_t)m_t + d_t - (1-\theta_t)C^2 - \theta_t l_t^+, 0)\\ &\quad + V_{t+1}(w_t^- + y_t + (1-\theta_t)m_t + d_t - (1-\theta_t)C^2 - \theta_t l_t^-, 0) \end{aligned} \tag{11}$$

Let $x_1 = w_t^+ + y_t + (1-\theta_t)m_t + d_t - (1-\theta_t)C^2 - \theta_t l_t^-$, $x_2 = w_t^+ + y_t + (1-\theta_t)m_t + d_t - (1-\theta_t)C^2 - \theta_t l_t^+$, $x_3 = w_t^- + y_t + (1-\theta_t)m_t + d_t - (1-\theta_t)C^2 - \theta_t l_t^-$ and $x_4 = w_t^- + y_t + (1-\theta_t)m_t + d_t - (1-\theta_t)C^2 - \theta_t l_t^+$. According to the jointly convexity of $V_t(w_t, l_t, m_t)$ in its arguments (see Lemma 4.3), we have

$$\begin{aligned} &V_{t+1}(x_2, 0) - V_{t+1}(x_1, 0) - V_{t+1}(x_4, 0) + V_{t+1}(x_3, 0)\\ &\quad \leq V_{t+1}(x_4, 0) - V_{t+1}(x_3, 0) - V_{t+1}(x_4, 0) + V_{t+1}(x_3, 0) = 0 \end{aligned} \tag{12}$$

As $V_{t+1}(w_{t+1}, l_{t+1}, m_{t+1})$ is assumed to be submodular in w_{t+1} and l_{t+1}, we have

$$V_{t+1}(x_2, 0) - V_{t+1}(x_1, 0) - V_{t+1}(x_2, 0) + V_{t+1}(x_1, 0) = 0 \tag{13}$$

According to the above analysis, we define w_{t+1} as x and we know $l_{t+1} = 0$. Equation (12) indicates that larger w_{t+1} increases the marginal cost of V_{t+1} when l_{t+1} is constant. Equation (13) indicates that larger w_{t+1} realizes for the increasing of w_t and the decreasing of l_t, which sheds light on the fact that the marginal cost of V_{t+1} caused by the increasing of w_t increases as l_t decreases. Evidently, $V_{t+1}(w_{t+1}, l_{t+1}, m_{t+1})$ is submodular in w_t and l_t.

For (i)(b) with any $w_t^+ \geq w_t^-$ and $l_t^+ \geq l_t^-$, we have

$$\begin{aligned} & V_{t+1}(w_t^+ - q_t + y_t, (1-\theta_t)C^2 + \theta_t l_t^+ - (1-\theta_t)m_t - d_t - q_t) \\ & \quad - V_{t+1}(w_t^+ - q_t + y_t, (1-\theta_t)C^2 + \theta_t l_t^- - (1-\theta_t)m_t - d_t - q_t) \\ & \quad - V_{t+1}(w_t^- - q_t + y_t, (1-\theta_t)C^2 + \theta_t l_t^+ - (1-\theta_t)m_t - d_t - q_t) \\ & \quad + V_{t+1}(w_t^- - q_t + y_t, (1-\theta_t)C^2 + \theta_t l_t^- - (1-\theta_t)m_t - d_t - q_t) \end{aligned} \tag{14}$$

Let $x_1 = w_t^+ - q_t + y_t$, $x_2 = w_t^- - q_t + y_t$, $y_1 = (1-\theta_t)C^2 + \theta_t l_t^+ - (1-\theta_t)m_t - d_t - q_t$ and $y_2 = (1-\theta_t)C^2 + \theta_t l_t^- - (1-\theta_t)m_t - d_t - q_t$. As $V_{t+1}(w_{t+1}, l_{t+1}, m_{t+1})$ is assumed to be submodular in w_{t+1} and l_{t+1}, we have

$$V_{t+1}(x_1, y_1) - V_{t+1}(x_1, y_2) - V_{t+1}(x_2, y_1) + V_{t+1}(x_2, y_2) \leq 0 \tag{15}$$

Equation (15) sheds light on the fact that the marginal cost of V_{t+1} caused by the increasing of w_{t+1} increases as l_{t+1} decreases. Since larger w_{t+1} realizes for the increasing of w_t and smaller l_{t+1} realizes for the decreasing of l_t, which shows that the marginal cost of V_{t+1} caused by the increasing of w_t increases as l_t decreases. Evidently, $V_{t+1}(w_{t+1}, l_{t+1}, m_{t+1})$ is submodular in w_t and l_t.

For (ii) with any $w_t^+ \geq w_t^-$ and $l_t^+ \geq l_t^-$, it is easy to prove that $V_{t+1}(w_t - \xi_t q_t + y_t, 0)$ is submodular in w_t and l_t, which completes the proof of (3).

(d) For any $d_t, y_t, \theta_t, \xi_t$, $V_t(w_t, l_t, m_t)$ is obtained by minimizing a convex function $U_t(q_t, w_t, l_t, m_t)$ in a convex feasible region over all q_t. Thus, we have the submodularity of $V_t(w_t, l_t, m_t)$ in (w_t, l_t). ■

The submodularity results derived above are crucial in characterizing the monotonicity of the optimal decisions in the system state variables. The monotonicity properties guide decision-makers with easy directions for policy adjustment with the change of system state.

Let $q_t^{*\max}(w_t, l_t, m_t)$ and $q_t^{*\min}(w_t, l_t, m_t)$ be the maximum and minimum optimal decision in epoch t for any given (w_t, l_t, m_t), respectively. Then we have the following results:

Corollary 4.5: *For every epoch t with any d_t, y_t, θ_t, ξ_t, the maximum and minimum optimal number of waitlisted elective patients who are scheduled to be admitted at stage 0, namely $q_t^{*\max}(w_t, l_t, m_t)$ and $q_t^{*\min}(w_t, l_t, m_t)$, respectively, are both increasing in the state (w_t, l_t).*

Proof: This directly follows from Theorem 4.4 and the fact that the feasible region is $\{(q_t, w_t, l_t) \in \mathbb{R}^3_+ | 0 \leq q_t \leq w_t\}$.

The monotonicity of the optimal decisions in the system state is quite intuitive. As mentioned above, q_t represents the number of waitlisted elective patients who are scheduled to be admitted in epoch t, w_t and l_t are state variables whose values are captured before decision q_t is made, correspondingly representing the number of backlogged elective patients on the waiting list and the number of unoccupied beds in the care unit. The monotonicity indicates that, either the number of elective hospitalization demand at upstream stage 0 increases or the number of available bed capacity at downstream stage 2 increases, the optimal number of waitlisted elective patients who are scheduled to be admitted increases. ■

5. Structural properties with random capacity

Above analyses assumed that the capacity vector (C^1, C^2) is fixed as constant. In this section, we transform the capacity of OR open time on any given day (C^1) and the total number of licensed beds in the care unit (C^2) into additional decision variables. Then we study how the optimal cost function V_t and the optimal decisions $q_t^{*\max}$ and $q_t^{*\min}$ change with respect to changes in (C^1, C^2). Capacity of OR time and hospitalization bed can vary in the practice of hospitals. Since the periodicity, seasonality, special holidays can impact on admission demand of certain disease types, for examples, there will be a large number of student ophthalmic surgeries in the summer holiday, many patients with digestive diseases in the Spring Festival, and a surge in respiratory diseases in autumn and winter, the hospital will adjust the total number of beds and the total open time of OR of the corresponding care units, with additional staff put to work in OR and non-licensed extra beds placed in ward corridors during peak demand seasons.

In the following, we first discuss the impact of capacity changes on the optimal cost V_t. According to the third and fourth terms of Equation (3), it is complex to analyse the joint convexity of $\hat{C}_t(q_t, w_t, l_t, m_t, C)$ in (q_t, C^1). Hence, for technical tractability, we simplify the third and fourth terms of Equation (3), i.e. $o[S(m_t + \xi_t q_t + d_t) - C^1]^+$ and $c[C^1 - S(m_t + \xi_t q_t + d_t)]^+$. Instead of assuming the OR capacity used by each patient is i.i.d., we assume that the total amount of OR capacity used by k patients on any given day is defined as uk, where u is a nonnegative random variable representing the average OR time used by a patient at stage 1. The OR time consumed by k patients represented by $S(k)$ and uk will not change the property of optimal scheduling policy and this simplification follows Liu et al. [1]. This approximation transforms the uncertainty of patient's surgery time into the uncertainty of the total open time of OR, based on which structural insights can be derived.

With these modifications, the model becomes

$$\begin{aligned}\hat{C}_t(q_t, w_t, l_t, m_t, C) = {} & b(w_t - q_t) + \gamma[(1-\xi_t)q_t - x_t]^+ \\ & + o[u(m_t + \xi_t q_t + d_t) - C^1]^+ \\ & + c[C^1 - u(m_t + \xi_t q_t + d_t)]^+ \\ & + p[\theta_t(C^2 - l_t - m_t) + m_t + \xi_t q_t + d_t - C^2]^+ \\ & + h[C^2 - \theta_t(C^2 - l_t - m_t) - m_t - q_t - d_t]^+ \qquad (16)\end{aligned}$$

$$C_t(q_t, w_t, l_t, m_t, C) = \mathbb{E}[\hat{C}_t(q_t, w_t, l_t, m_t, C)] \qquad (17)$$

$$\begin{aligned}U_t(q_t, w_t, l_t, m_t, C) = {} & C_t(q_t, w_t, l_t, m_t, C) \\ & + \alpha\mathbb{E}[V_{t+1}(w_{t+1}, l_{t+1}, m_{t+1}, C)|(w_t, l_t, m_t)] \qquad (18)\end{aligned}$$

$$V_t(w_t, l_t, m_t, C) = \min_{0 \le q_t \le w_t} U_t(q_t, w_t, l_t, m_t, C) \qquad (19)$$

where $C = (C^1, C^2)$. Then we can show the following results similar to Lemmas 4.1– 4.3.

Theorem 5.1: *For every epoch t with any $d_t, y_t, \theta_t, \xi_t$, $C_t(q_t, w_t, l_t, m_t, C)$ is jointly convex in its arguments.*

Proof: According to Equation (17), we discuss the joint convexity of it in two cases.

First, when $x_t = (C^2 - \theta_t(C^2 - l_t - m_t) - m_t - d_t - \xi_t q_t)^+ > 0$ holds, we have

$$\begin{aligned}& \hat{C}_t(q_t, w_t, l_t, m_t, C) \\ & = b(w_t - q_t) + \gamma[(1-\theta_t)m_t + d_t + q_t - (1-\theta_t)C^2 - \theta_t l_t]^+ \\ & \quad + o[u(m_t + \xi_t q_t + d_t) - C^1]^+ + c[C^1 - u(m_t + \xi_t q_t + d_t)]^+ \\ & \quad + h[C^2 - \theta_t(C^2 - l_t - m_t) - m_t - q_t - d_t]^+ \qquad (20)\end{aligned}$$

Second, when $x_t = (C^2 - \theta_t(C^2 - l_t - m_t) - m_t - d_t - \xi_t q_t)^+ = 0$ holds, i.e. $C^2 - \theta_t(C^2 - l_t - m_t) - m_t - d_t - \xi_t q_t \le 0$, we have

$$\begin{aligned}& \hat{C}_t(q_t, w_t, l_t, m_t, C) \\ & = b(w_t - q_t) + \gamma(1-\xi_t)q_t \\ & \quad + o[u(m_t + \xi_t q_t + d_t) - C^1]^+ + c[C^1 - u(m_t + \xi_t q_t + d_t)]^+ \\ & \quad + p[\theta_t(C^2 - l_t - m_t) + m_t + \xi_t q_t + d_t - C^2] \qquad (21)\end{aligned}$$

According to Equations (20) and (21), $\hat{C}_t(q_t, w_t, l_t, m_t, C)$ is evidently joint convex in its arguments (q_t, w_t, l_t, m_t, C), and thus $C_t(q_t, w_t, l_t, m_t, C)$ is joint convex based on Equation (17). The rest of the proof is similar to that of Lemma 4.1. ■

Theorem 5.2: *For every epoch t with any d_t, y_t, θ_t, ξ_t, w_{t+1} and l_{t+1} are jointly convex in their arguments, respectively.*

Proof: According to Equation (1), i.e. $w_{t+1} = w_t - q_t + y_t + [(1-\xi_t)q_t - x_t]^+$, and Equation (2), i.e. $l_{t+1} = x_t - (1-\xi_t)q_t + [(1-\xi_t)q_t - x_t]^+$, we discuss the joint convexity of w_{t+1} and l_{t+1} in two cases.

First, when $x_t = (C^2 - \theta_t(C^2 - l_t - m_t) - m_t - d_t - \xi_t q_t)^+ > 0$ holds, we have

$$\begin{aligned} w_{t+1} &= w_t - q_t + y_t + [(1-\theta_t)m_t + d_t + q_t - (1-\theta_t)C^2 - \theta_t l_t]^+ \\ l_{t+1} &= (1-\theta_t)C^2 + \theta_t l_t - (1-\theta_t)m_t - d_t - q_t + ((1-\theta_t)m_t + d_t \\ &\quad + q_t - (1-\theta_t)C^2 - \theta_t l_t)^+ \end{aligned} \tag{22}$$

Second, when $x_t = (C^2 - \theta_t(C^2 - l_t - m_t) - m_t - d_t - \xi_t q_t)^+ = 0$ holds, we have

$$\begin{aligned} w_{t+1} &= w_t - \xi_t q_t + y_t \\ l_{t+1} &= 0 \end{aligned} \tag{23}$$

According to Equations (22) and (23), it is easy to check the joint convexity of w_{t+1} and l_{t+1} in their arguments, respectively. ■

Theorem 5.3: *For every epoch t with any d_t, y_t, θ_t, ξ_t, $U_t(q_t, w_t, l_t, m_t, C)$ and $V_t(w_t, l_t, m_t, C)$ are jointly convex in their arguments, respectively.*

Proof: According to Theorems 5.1 and 5.2, the proof is similar to that of Lemma 4.3. Assuming that $V_{t+1}(w_{t+1}, l_{t+1}, m_{t+1}, C)$ is jointly convex in its arguments (which is true for $t = T$). We know that w_{t+1} and l_{t+1} are jointly convex in their arguments (q_t, w_t, l_t, m_t, C), respectively. Hence, $\mathbb{E}[V_{t+1}(w_{t+1}, l_{t+1}, m_{t+1})]$ is jointly convex in its arguments. Therefore, $U_t(q_t, w_t, l_t, m_t, C)$ is jointly convex in its arguments. According to Equation (19), $V_t(w_t, l_t, m_t, C)$ is obtained by minimizing a convex function $U_t(q_t, w_t, l_t, m_t, C)$ in a convex feasible region, it follows that $V_t(w_t, l_t, m_t, C)$ is also convex (see Theorem A.4 in [31]). By induction, the property holds for all t.

The theorems above suggests a diminishing return as the capacity of OR time and bed increases. In the following section, we study how the optimal decisions change with varying levels of capacity. For the same technical reason above, we still assume that the total amount of OR capacity used by k patients on any given day is defined as uk, where u is a non-negative random variable representing the average OR time used by a patient at stage 1. ■

Corollary 5.4: *For every epoch t with any d_t, y_t, θ_t, ξ_t, and every fixed C^1, the following properties $C_t(q_t, w_t, l_t, m_t, C)$ (17), $U_t(q_t, w_t, l_t, m_t, C)$ (18) and $V_t(w_t, l_t, m_t, C)$ (19) hold:*

(1) *$C_t(q_t, w_t, l_t, m_t, C)$ is submodular in (q_t, w_t, l_t, C^2).*
(2) *$U_t(q_t, w_t, l_t, m_t, C)$ is submodular in (q_t, w_t, l_t, C^2).*
(3) *$V_t(w_t, l_t, m_t, C)$ is submodular in (w_t, l_t, C^2).*

Proof: The proof is similar to that of Theorem 4.4, and thus we omit it here. The submodularity implies the marginal cost caused by the investment of bed capacity decreases with the increasing of the number of waitlisted elective patients who are scheduled to be admitted, which characterizes the monotonicity of the scheduling decisions in the level of bed capacity (see the corollary formalized below). ■

Corollary 5.5: *For every epoch t with any d_t, y_t, θ_t, ξ_t, m_t, and every fixed C^1, the maximum and minimum optimal number of waitlisted elective patients who are scheduled to be admitted at stage 0, namely $q_t^{*\max}(w_t, l_t, C)$ and $q_t^{*\min}(w_t, l_t, C)$, respectively, are both increasing in the level of bed capacity C^2.*

The monotonicity property stated in the corollary is easy to give an explanation. Any increase of bed capacity enables more patients to be accommodated in the system regardless of the current level of OR capacity. However, the result does not hold for any increase of OR capacity. That is, if we fix the bed capacity level C^2 and increase the OR capacity level C^1, the optimal decisions $q_t^{*\max}(w_t, l_t, C)$ and $q_t^{*\min}(w_t, l_t, C)$ might not increase. Since the average length of surgery is assumed to be constant, the number of surgeries increases as the total open time of OR, which directly affects the m_t number of scheduled elective surgeries before the admission decision q_t is made. After bed dispatcher is notified of m_t, (s)he decides q_t considering the balance of costs incurred at three stages. Hence, the increase of OR capacity might result in the increase of m_t while the decision of q_t stays unaffected.

6. Numerical studies

As discussed above, this paper is the first to analytically study dynamic admission scheduling problem with considering the capacity usage of two inter-related resources (beds and operating rooms) between three consecutive stages of care during surgical patients' admissions to Chinese public hospitals. By formulating a stochastic dynamic programming model with a single decision variable and multiple state variables, the optimal scheduling policy is derived and characterized. The monotonicity property of the optimal decisions in the system state guides decision-makers with easy directions for policy adjustment with the change of system state.

We have been working with the West China Hospital (WCH) at Sichuan University, one of the largest hospitals in China. This hospital operates a large inpatient department with a capacity of 4300 licensed beds shared by 44 specialty care units, including 23 non-surgical units and 21 surgical units. It is a public hospital to serve the general population, but it also must be financially self-sufficient. Like most other tertiary hospitals in this region, it has been overcrowded (5.44 million outpatients and emergency patients, 263,700 admissions, and 175,300 surgeries in 2018) and waiting lists for hospitalization have been growing. Since the average daily discharge volume of the entire hospital is nearly 600 people, inpatient beds are rather limited and cannot satisfy all demand in a timely manner. Waiting times for elective patients varied from several weeks to more than 1 year. Here, we use process flows and data from the WCH, the numerical experiments in this section follow up on the theoretical work above to investigate the performance of our proposed scheduling method and compare it with scheduling policies that make decisions independently of operations in other units. All numerical experiments were conducted and solved with the Optimization Toolbox package of MATLAB R2015a. All computations were run on a MacOS X EI Capitan (Version 10.11.6) Intel Core i5 CPU with 2.7 gigahertz.

6.1. Optimal decisions

According to Corollary 4.5, for every epoch t with any $d_t, y_t, \theta_t, \xi_t$, the maximum and minimum optimal number of waitlisted elective patients who are scheduled to be admitted at stage 0, namely $q_t^{*\max}(w_t, l_t, m_t)$ and $q_t^{*\min}(w_t, l_t, m_t)$, respectively, are both increasing in the state (w_t, l_t). The monotonicity indicates that, either the number of elective hospitalization demand at upstream stage 0 increases or the number of available bed capacity at downstream stage 2 increases, the optimal number of waitlisted elective patients who are scheduled to be admitted increases.

We conduct our numerical experiments with data collected from the USCU at WCH. Raw data cover the period of January to October 2015 (only workdays recorded) and include three sets: (1) all admission observations (3507), (2) all cancellation records (540), and (3) all requests that are still waiting for beds (260). The waiting list data as the input to the multi-period models are generated by integrating these three sets of raw data recorded daily at the patient level. By analyzing the data, the results show that the average number of daily discharged inpatients is 22. Since the total number of beds at USCU is 140, the average discharge rate is about 0.15. Hence, $E[\theta_t] = 0.85$ holds. Moreover, we have $E[\xi_t] = 0.5$, $E[m_t] = 3$, $E[d_t] = 1.5$, $E[y_t] = 8$, $E[u] = 1$, and the initial state is $w_0 = 18$ and $l_0 = 13$. According to field observation and consultation with bed manager and corresponding planners, USCU is a large care unit with multiple wards. The total bed capacity of USCU results in a huge state space of the dynamic system, commonly called the curse of dimensionality. For technical tractability, we consider a single OR with capacity $C^1 = 8$ and a single ward with bed capacity $C^2 = 30$. Except for the curse of dimensionality problem of stochastic dynamic programming models, the reason why we consider a single OR with 8 hours open time and a single ward with 30 beds is that, all 140 beds of USCU have been allocated to different medical groups that each has a leader physician and these medical groups can only use their own beds to admit patients. For example, group A has 30 beds; group B has 40 beds, and so on. In practice, each ward at WCH has one to more than six beds, depending on the type of specialty care unit. Hence, the essence of a ward with 30 beds means a medical group with several wards including a total number of 30 beds. Since each patient's admission certification is signed by the patient's own doctor during the OD or ED service stage (see details from the Section 3.1), each patient is waiting for a bed of his doctor's medical group. Several wards with a total number of 30 beds and a ward with 30 beds will not affect our model and the results. Moreover, WCH implements a master surgical schedule policy and each medical group has its own fixed surgery day and operating theatre. The normal open time of an operating theatre at WCH is 8 h, and each patient is waiting for his/her doctor's surgery day. Every day the bed scheduler who is in charge of the admission scheduling and resource allocation of USCU make decisions of different medical groups independently considering their available bed and OR capacity. Hence, it is reasonable to consider a single OR with a capacity of 8 hours and a ward with 30 beds in our study setting. The cost parameters are shown in Table 2. Based on above data and information, the optimal number of waitlisted elective patients who are scheduled to be admitted at stage 0, namely q^*, is obtained via MATLAB R2015a (see Table 3 and Figure 5) and the computation time is 17.4 minutes.

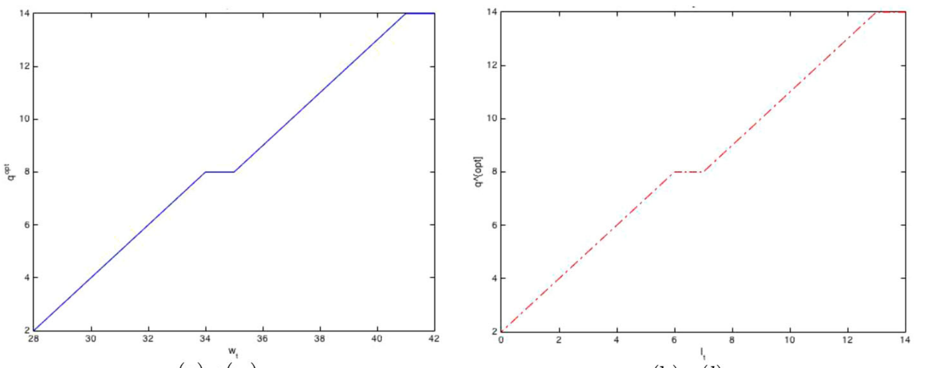
The results in Table 3 and Figure 5 indicate that for any given l_t and other variables, the optimal number of elective patients who are scheduled to be admitted increases as the number of waitlisted elective patients, when the optimal decision increases to a certain number it stays invariable with the increasing of the number of waitlisted elective patients. This implies that the optimal policy q_t^* is a control limit policy for w_t, and the increment of q_t^* is no more than one unit as w_t increases by one unit. The same conclusions hold for l_t. In addition, when w_t and l_t increase simultaneously for any given other variables, the optimal policy q_t^* first increases gradually and finally maintains the maximum optimal number ($q_t^{*\max}$) staying unchanged. The monotonicity of the optimal decisions in the system state is verified by numerical experiments.

Table 2. The cost parameters.

Parameter	b	γ	o	c	p	h	α
Value	1	15	10	10	0.5	0.2	0.8

Table 3. The optimal solution (q^*) when $t = 4$.

w_t	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42
l_t	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
q^*	2	3	4	5	6	7	8	8	9	10	11	12	13	14	14



(a) $q(w)$.

(b) $q(l)$.

Figure 5. Results of (a) and (b) describe that the optimal solution q^* is non-decreasing in both w_t and l_t.

6.2. Convexity of the value function V_t in the model with random capacity

This section presents numerical experiments on how the value function (V_t) and the optimal decisions (q_t^*) change with the capacity levels (C^1, C^2). For every epoch t with any $d_t, y_t, \theta_t, \xi_t$ and m_t, the numerical experiments are conducted to verify (a) when the bed capacity C^2 maintains unchanged, the value function V_t is convex in the OR capacity C^1; (b) when the OR capacity C^1 maintains unchanged, the value function V_t is convex in the bed capacity C^2; (c) for any given C^2, the optimal decision q_t^* may not increase in C^1. In addition, the optimal capacity of one stage depends on the relative weight of costs in all involved stages, the total system cost is more sensitive to the capacity change in a stage that carries higher costs. The cost parameters are shown in Table 4.

(a) The case when the bed capacity C^2 maintains unchanged. We have $C^2 = 30,\ C^1 \in [2, 12],\ T = 10,\ E[\theta_t] = 0.85,\ E[\xi_t] = 0.5,\ E[m_t] = 3,\ E[d_t] = 1.5, E[y_t] = 8, E[u] = 1$, and the initial state is $w_0 = 18$ and $l_0 = 13$. According to the theoretical results, the optimal value function V_0 and the optimal scheduling policy q_t^* is obtained via MATLAB R2015a. The results of V_0 are shown in Table 5 and Figure 6.

From Table 5 and Figure 6, we have:

(1) When the bed capacity C^2 maintains unchanged, the optimal value function $V_0(C^1, w_0, l_0)$ is convex in the OR capacity C^1, either with higher OR related costs or with higher bed-related costs. The results indicate that with other variables remain unchanged, the total discounted expected cost of the

Table 4. The cost parameters.

Parameter	b	γ	o	c	p	h	α
Higher OR related costs	1	15	10	10	0.5	0.2	0.8
Higher bed-related costs	1	15	1	1	3	1.2	0.8

Table 5. The results of $V_0(C^1, w_0, l_0)$ based on two cost parameter cases.

$o = c = 10, p = 0.5, h = 0.2$ $V_0(C^1, w_0, l_0)$	$C^1 = 2$ 259.8	$C^1 = 3$ 213.18	$C^1 = 4$ 168.6	$C^1 = 5$ 133.56	$C^1 = 6$ 113.5	$C^1 = 7$ 123.04
$o = c = 10, p = 0.5, h = 0.2$ $V_0(C^1, w_0, l_0)$	$C^1 = 8$ 143.01	$C^1 = 9$ 177.72	$C^1 = 10$ 214.22	$C^1 = 11$ 255.58	$C^1 = 12$ 301.88	/ /
$o = c = 1, p = 3, h = 1.2$ $V_0(C^1, w_0, l_0)$	$C^1 = 2$ 85.83	$C^1 = 3$ 84.64	$C^1 = 4$ 78.8	$C^1 = 5$ 76.36	$C^1 = 6$ 76.24	$C^1 = 7$ 70.69
$o = c = 1, p = 3, h = 1.2$ $V_0(C^1, w_0, l_0)$	$C^1 = 8$ 75.67	$C^1 = 9$ 78.32	$C^1 = 10$ 81.91	$C^1 = 11$ 84.16	$C^1 = 12$ 88.79	/ /

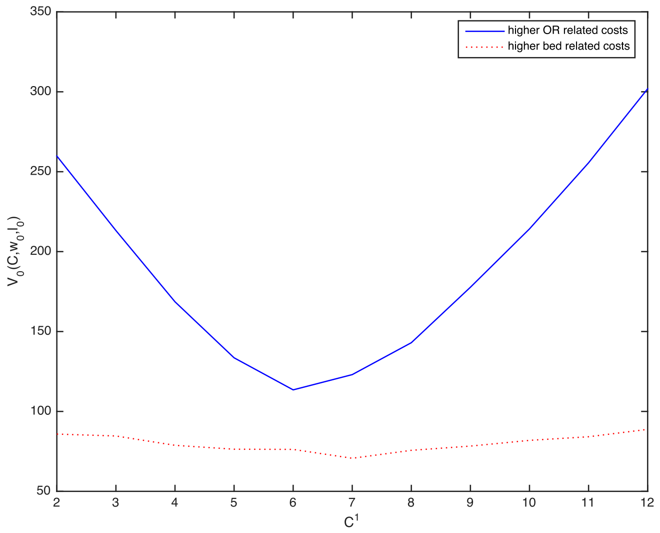


Figure 6. The value function $V_0(C, w_0, l_0)$ is convex in the OR capacity C^1.

system exhibits decreasing marginal returns as the OR capacity in stage 1 increases independently of other stages.

(2) With higher OR related costs, the total expected cost V_t is more sensitive to the varying of C^1.

(3) The optimal value of C^1 is sensitive to the relative weight between OR related costs (o, c) and bed-related costs (p, h). With higher OR related costs, the optimal value of C^1 is 6, the system throughput is mainly driven by the need to use the OR capacity efficiently. Otherwise, with higher bed-related costs, the optimal value of C^1 is 7, as the system throughput is determined by the bed capacity.

(b) The case when the OR capacity C^1 maintains unchanged. We have $C^1 = 8$, $C^2 \in [21, 40]$, $T = 10$, $E[\theta_t] = 0.85$, $E[\xi_t] = 0.5$, $E[m_t] = 3$, $E[d_t] = 1.5$, $E[y_t] = 8$, $E[u] = 1$, and the initial state is $w_0 = 18$ and $l_0 = 13$. According to the theoretical results, the optimal value function V_0 and the optimal scheduling policy q_t^* is obtained via MATLAB R2015a. The results of V_0 are shown in Table 6 and Figure 7.

Table 6. The results of $V_0(C^2, w_0, l_0)$ based on two cost parameter cases.

$o = c = 10, p = 0.5, h = 0.2$ $V_0(C^2, w_0, l_0)$	$C^2 = 21$ 182.73	$C^2 = 22$ 178.38	$C^2 = 23$ 172.61	$C^2 = 24$ 164.12	$C^2 = 25$ 159.72	$C^2 = 26$ 153.19	$C^2 = 27$ 150.61
$o = c = 10, p = 0.5, h = 0.2$ $V_0(C^1, w_0, l_0)$	$C^2 = 28$ 146.75	$C^2 = 29$ 146.20	$C^2 = 30$ 143.00	$C^2 = 31$ 140.31	$C^2 = 32$ 135.28	$C^2 = 33$ 131.99	$C^2 = 34$ 130.81
$o = c = 10, p = 0.5, h = 0.2$ $V_0(C^1, w_0, l_0)$	$C^2 = 35$ 127.29	$C^2 = 36$ 122.36	$C^2 = 37$ 123.30	$C^2 = 38$ 123.37	$C^2 = 39$ 124.51	$C^2 = 40$ 129.40	/ /
$o = c = 1, p = 3, h = 1.2$ $V_0(C^2, w_0, l_0)$	$C^2 = 21$ 130.56	$C^2 = 22$ 119.37	$C^2 = 23$ 113.41	$C^2 = 24$ 107.92	$C^2 = 25$ 101.27	$C^2 = 26$ 93.73	$C^2 = 27$ 90.13
$o = c = 1, p = 3, h = 1.2$ $V_0(C^1, w_0, l_0)$	$C^2 = 28$ 85.04	$C^2 = 29$ 80.25	$C^2 = 30$ 79.19	$C^2 = 31$ 78.92	$C^2 = 32$ 81.67	$C^2 = 33$ 84.54	$C^2 = 34$ 87.87
$o = c = 1, p = 3, h = 1.2$ $V_0(C^1, w_0, l_0)$	$C^2 = 35$ 90.83	$C^2 = 36$ 93.13	$C^2 = 37$ 97.32	$C^2 = 38$ 101.23	$C^2 = 39$ 105.77	$C^2 = 40$ 110.69	/ /

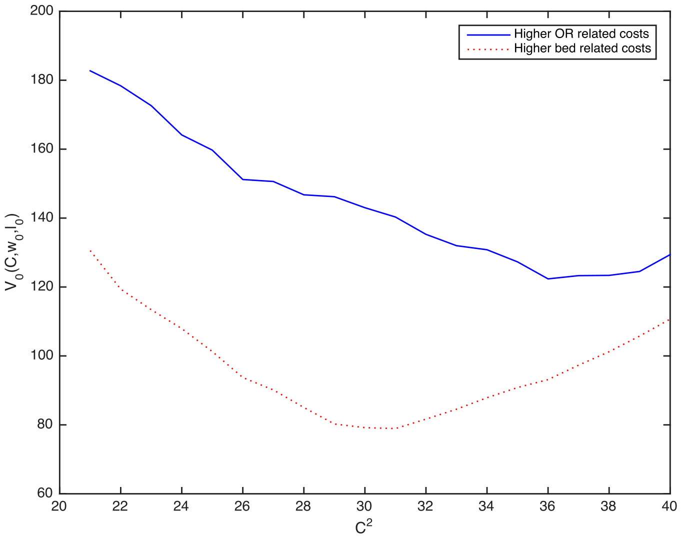


Figure 7. The value function $V_0(C, w_0, l_0)$ is convex in the bed capacity C^2.

From Table 6 and Figure 7, we have:

(1) When the OR capacity C^1 maintains unchanged, the optimal value function $V_0(C^2, w_0, l_0)$ is convex in the bed capacity C^2, either with higher OR related costs or with higher bed-related costs. The results indicate that with other variables remain unchanged, the total discounted expected cost of the system exhibits decreasing marginal returns as the bed capacity increase independently of stage 1.

(2) With higher bed-related costs, the total expected cost V_t is more sensitive to the varying of C^2.

(3) The optimal value of C^2 is sensitive to the relative weight between OR related costs (o, c) and bed-related costs (p, h). With higher OR related costs, the optimal value of C^2 is 36, the system throughput is mainly driven by the need to use the OR capacity efficiently. Otherwise, with higher bed-related

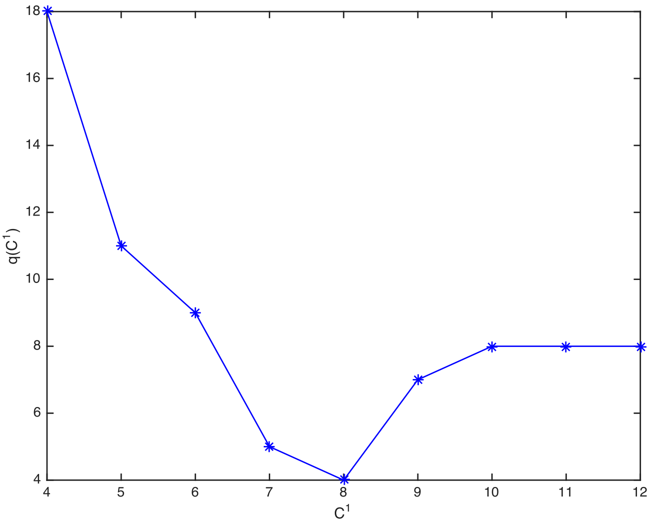


Figure 8. The optimal policy $q(C, w_0, l_0)$ is not monotonic in the OR capacity C^1.

costs, the optimal value of C^2 is 31, as the system throughput is determined by the bed capacity.

(c) The monotonicity of the optimal scheduling policy q_t^* in the OR capacity C^1. We have $C^2 = 30, C^1 \in [4, 12], T = 10, E[\theta_t] = 0.85, E[\xi_t] = 0.5, E[m_t] = 3$, $E[d_t] = 1.5, E[y_t] = 8, E[u] = 1$, and the initial state is $w_0 = 18$ and $l_0 = 13$. We assume cost parameters are $b = 1, \gamma = 15, o = c = 5, p = 2, h = 1.2$. The varying pattern of q_t^* in C^1 is shown in Figure 8, which indicates that the optimal scheduling policy q_t^* is not monotonic in the OR capacity C^1. When $C^1 = 4$, the number of waitlisted elective patients is 18, the number of scheduled elective surgeries is 3, in this case, the OR capacity is seriously insufficient, incurring OR overtime cost. Since the OR overtime and idling costs are both five times of the waiting cost and exceed bed shortage and idling costs, to avoid excessive OR overtime cost, the optimal scheduling policy would make patients wait on the list, namely the optimal number of elective patients admitted decreases. Then with the increase of OR capacity, the admitted elective patients occupy OR capacity without incurring overtime cost. However, when the OR capacity increases to a certain number (i.e. $C^1 = 8$), the optimal scheduling policy turns to control the OR idling cost instead of avoiding excessive overtime cost, in this case, the optimal number of elective patients admitted increases.

6.3. Comparisons of the optimal policy based on joint scheduling and isolated scheduling

This section presents the comparisons of the optimal policy based on joint scheduling and isolated scheduling, to clarify the value of making joint scheduling decisions compared to isolated scheduling decisions. Intuitively, we test the empirical performance of the following three kind of policies:

(a) The optimal policy (Opt) based on joint scheduling: in this policy, variables q_t, w_t, and l_t are all discrete non-negative integers;
(b) The OR single-resource scheduling policy (Sin1): in this policy, the total bed capacity is assumed to be infinity, the optimal scheduling policy only considers the constraint of OR capacity;
(c) The bed single-resource scheduling policy (Sin2): in this policy, the total OR capacity is assumed to be infinity, the optimal scheduling policy only considers the constraint of bed capacity.

The single-resource scheduling policy only considers a capacity constraint of this specific unit, without considering the influence of this scheduling policy on the usage of capacity in other units. Currently, bed dispatchers of WCH do not consider the constraint of OR capacity when making admission scheduling decisions, and the surgery scheduling does not consider the usage of bed capacity in the downstream stage. Subsequently, we compare the performance of the optimal policy (Opt) against two single-resource policies (Sin1 and Sin2). The two isolated scheduling policies reflect the actual setting of the study hospital. A large performance gap between the optimal policy and the single-resource policies indicate that joint scheduling is valuable in multi-unit multi-resource scheduling.

First, we compare the performance of Opt and Sin1, where we chose $C^1 = 8$ and the bed capacity C^2 is varied in the tests. As above mentioned, we have $T = 10$, $E[\theta_t] = 0.85$, $E[\xi_t] = 0.5$, $E[m_t] = 3$, $E[d_t] = 1.5$, $E[y_t] = 8$, $E[u] = 1$, and the initial state is $w_0 = 18$ and $l_0 = 13$. The waiting cost is normalized at $b = 1$, and the over-scheduled cost is $\gamma = 15$. For each combination of the chosen cost parameters o, c, p, h and C^2, we calculate the total discounted expected costs under each of the scheduling policies, i.e. Opt and Sin1. For easy comparison, we present all results as the performance ratio with respect to the total discounted cost under Opt. The ratios $V_0^{\text{Sin1}}/V_0^{\text{Opt}}$ are summarized in Table 7.

When the total bed capacity is assumed to be infinity under Sin1, the cost ratios indicate the system performance following the decisions of a surgery scheduler who does not pay attention to the bed capacity usage. Higher ratios correspond to more cost saving when joint scheduling is conducted. From Table 7, when $o = c = 10$, $p = 0.5$, $h = 0.2$, the average ratio is 2488%; when $o = c = 10$, $p = 30, h = 10$, the average ratio is 2099%; when $o = 10, c = 1, p = 30, h = 15$, the average ratio is 414%, suggesting that the system overspends four-fold when not making joint scheduling under the best cost parameter scenario. More importantly, as the ratio of OR related costs with respect to bed-related costs increases gradually, i.e. varying from $o = 10, c = 1, p = 30, h = 15$ to $o = c = 10, p = 30$, $h = 10$ and $o = c = 10, p = 0.5, h = 0.2$, we observe that the cost ratios increase

Table 7. The comparison results of Opt and Sin1.

$V_0^{\mathrm{Sin1}}/V_0^{\mathrm{Opt}}$	$o = c = 10, p = 0.5, h = 0.2$	$o = c = 10, p = 30, h = 10$	$o = 10, c = 1, p = 30, h = 15$
$C^2 = 21$	20.793	18.048	3.869
$C^2 = 22$	20.793	18.048	3.869
$C^2 = 23$	22.055	18.569	3.723
$C^2 = 24$	21.541	18.899	3.816
$C^2 = 25$	22.630	19.669	4.034
$C^2 = 26$	23.306	20.551	4.289
$C^2 = 27$	23.579	20.301	4.118
$C^2 = 28$	23.541	19.922	3.939
$C^2 = 29$	23.632	20.160	3.854
$C^2 = 30$	25.465	21.885	4.155
$C^2 = 31$	27.065	22.142	4.200
$C^2 = 32$	28.434	23.687	4.540
$C^2 = 33$	29.535	25.047	4.905
$C^2 = 34$	30.366	24.581	4.640
$C^2 = 35$	30.679	23.596	4.340

under each value of C^2, indicating that Sin1, in general, performs better when the shortage cost of OR capacity is smaller. This is because the total bed capacity is assumed to be infinity under Sin1, OR capacity becomes the bottleneck resource, making the total discounted expected cost under Sin1 more sensitive to OR related costs. The explanation is that when the OR related costs become more significant, a policy that aims to minimize costs in the OR stage is likely to perform well.

Subsequently, we compare the performance of Opt and Sin2, where we chose $C^2 = 30$ and the OR capacity C^1 is varied in the tests. As above mentioned, we have $T = 10$, $E[\theta_t] = 0.85$, $E[\xi_t] = 0.5$, $E[m_t] = 3$, $E[d_t] = 1.5$, $E[y_t] = 8$, $E[u] = 1$, and the initial state is $w_0 = 18$ and $l_0 = 13$. The waiting cost is normalized at $b = 1$, and the over-scheduled cost is $\gamma = 15$. For each combination of the chosen cost parameters o, c, p, h and C^1, we calculate the total discounted expected costs under each of the scheduling policies, i.e. Opt and Sin2. For easy comparison, we present all results as the performance ratio with respect to the total discounted cost under Opt. The ratios $V_0^{\mathrm{Sin2}}/V_0^{\mathrm{Opt}}$ are summarized in Table 8.

Table 8. The comparison results of Opt and Sin2.

$V_0^{\mathrm{Sin2}}/V_0^{\mathrm{Opt}}$	$o = c = 10, p = h = 5$	$o = c = 10, p = 30, h = 10$	$o = 10, c = 1, p = 30, h = 15$
$C^1 = 2$	6.930	12.296	17.319
$C^1 = 3$	8.063	14.370	20.203
$C^1 = 4$	9.731	17.378	24.326
$C^1 = 5$	12.432	22.131	30.701
$C^1 = 6$	13.849	26.592	39.335
$C^1 = 7$	10.773	20.638	39.365
$C^1 = 8$	8.839	16.786	39.406
$C^1 = 9$	7.440	14.248	39.610
$C^1 = 10$	6.459	12.478	39.948
$C^1 = 11$	5.767	11.157	40.347
$C^1 = 12$	5.205	9.954	39.801

When the total OR capacity is assumed to be infinity under Sin2, the cost ratios represent the performance of a system admitting patients according to an admission scheduler who ignores the operations in the OR unit. From Table 8, when $o = c = 10, p = h = 5$, the average ratio is 868%, suggesting that the system overspends eight-fold when not making joint scheduling; when $o = c = 10$, $p = 30, h = 10$, the average ratio is 1618%; when $o = 10, c = 1, p = 30, h = 15$, the average ratio is 3367%. More importantly, as the OR capacity C^1 gets larger, we observe that the cost ratios first increase then decrease under each combination of cost parameters. When C^1 is relatively small, the OR capacity is, in fact, the bottleneck resource of Opt, whereas policy Sin2 assumes infinite OR capacity, thus the total discounted cost under Sin2 is relatively slightly higher than the optimal cost under Opt. As C^1 becomes larger, bed becomes the bottleneck resource for policy Sin2, the value of dynamically balancing the usage of capacity of multiple resources becomes more significant, and as a result we observe that the performance gap between Sin2 and Opt increases as C^1 increases. Interestingly, however, we find that after C^1 is raised to a certain level, the performance gap of Sin2 and Opt starts to become small as C^1 increases, indicating that the bottleneck resource of Opt becomes the bed capacity. Moreover, as the ratio of bed-related costs with respect to OR related costs increases gradually, i.e. varying from $o = c = 10, p = h = 5$ to $o = c = 10, p = 30, h = 10$ and $o = 10, c = 1$, $p = 30, h = 15$, we observe that the cost ratios increase under each value of C^1, indicating that Sin2, in general, performs better when the shortage cost of bed capacity is smaller.

In summary, policies Sin1 and Sin2, which make scheduling decisions based on capacity usage of only one resource in the system, can lead to significant inefficiency and financial loss. However, they can perform better either when the bottleneck resource is correctly identified and has much smaller capacity than the other resource, or when the related costs of the bottleneck resource are lower than those of the other resource. Conversely, the joint scheduling policy can bring substantial cost savings to the system. Even by identifying the bottleneck resource correctly, joint scheduling is still most beneficial against isolated scheduling policies like Sin1 and Sin2.

7. Discussions and conclusions

Public hospital managers are under tremendous pressure to balance the use of capacity in different units to ensure smooth patient flow through the whole medical service system. Suboptimal decisions are often drawn because many hospital units still operate in isolation ignoring the influences of other services or the impact of the change on the overall care chain, which often leads to poor scheduling, blocking, inefficient use of capacity, and consequently, high cost and reduced quality of care. As the basis of the National Healthcare System, public hospitals in China are self-financing institutions in China's market economy system and have profit incentives. Hence, the optimization of joint scheduling decisions on capacity usage in multiple units of a hospital is important and challenging for hospital managers.

In this paper, based on a dynamic multi-unit multi-stage admission control system, a joint scheduling decision model based on integrated capacity usage of the hospital bed and the operating room is studied from the perspective of cost minimization of the whole system. The decision on the number of elective patients to be admitted on a given day is determined to balance the costs of capacity over-utilization and under-utilization with the cost of making patients wait. Optimal policies derived from using the model enable hospitals to substantially improve their operational efficiency and effectiveness compared to policies from using isolated scheduling decisions that only focus on a single unit of the hospital. First, the model systematically formulates the uncertainty, dynamic and linkage of capacity usage of OR and bed, which corresponds to the practice environment of the use of healthcare capacity after China's health care reform. Second, the model breaks the limitations of traditional single-unit isolated scheduling decision rules, which copes with the problem of system global optimization. Third, starting from two conditions of fixed capacity and variable capacity, how the optimal policies under the joint scheduling decision rule change with the varying of system state and capacity of OR and bed is characterized to optimize the cost of the whole system.

Our work makes an important contribution by firstly formulating and analyzing a dynamic multi-unit and multi-resource scheduling model that integrates information about capacity usage at more than two service stages in hospitals. Particularly, we characterize the monotonicity of the number of elective admits in each period in the system state and in the bed capacity, indicating that a higher level of waiting elective patients and available (or total) bed capacity pulls more elective patients through the system, thereby providing useful guidelines for adjusting scheduling decisions in practice. In addition, we show that the total discounted expected cost of the system exhibits decreasing marginal returns as the capacity in each stage increases independently of the other stage, which provides intuitive insights on capacity adjustment for practitioners to refer. Through numerical experiments based on realistic data from practice, we show that there is substantial value to making joint scheduling decisions compared to isolated scheduling decisions.

Our model has some limitations and can be extended in a variety of directions to formulate more details in reality. Specifically, the decision of elective admission control only considers capacity joint scheduling before the admission service day, whereas the scheduled elective patients' behaviours like *no-show* and *cancellation* that might occur on the service day are ignored without considering the real-time scheduling of capacity usage of OR and bed. This paper considers only one resource constraint in each stage, in fact, there are multiple resources used in each stage and each with its own over-utilization and under-utilization cost, such as OR, nurse, surgeon and anaesthesiologist in the OR stage. Particularly, for technical tractability, numerical experiments conducted for performance validation of our theoretical model considered a single OR with 8 hours open time and a single ward with 30 beds. Though the assumption is realistic, it simplified the actual problem of operational management of wards and ORs in many large hospitals, resulting in an admission scheduling policy that may not be robust enough. Thus, the generalizability and portability of the model and optimal policy in this study in other hospitals (especially those who are scheduling multiple wards and ORs simultaneously) is another possible extension. Such an extension would be a necessary next step before attempting to disseminate our model beyond this Chinese health system. Moreover, the decision in each period is made in the afternoon of the day before the admission service day in the current practice of Chinese public hospitals, however, this decision rule is inconvenient for patients and hospital managers to prepare for admission and adjust decisions, respectively. Hence, advanced scheduling that decision is made a few days (such as one week) before the admission service day is necessary to be studied. More importantly, public health emergencies like COVID-19 have brought greater challenges to the resource scheduling of large hospitals globally, emergency demand with infectious and fatal viruses and regular demand with severe diseases like cancers and chronic illnesses need many medical resources simultaneously, joint scheduling of doctors, nurses, ORs, beds, protection equipment and other resources would be more valuable in such setting.

In summary, this paper presents a useful model to guide joint scheduling decisions in a system with three consecutive service stages and two resource bottlenecks. Our work formulates the scheduling setting that is more complex than previous similar studies. Future research could focus on the above-mentioned directions that should be extended.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Department of Science and Technology of Sichuan Province [2019YFS0276] and the National Natural Science Foundation of China [71532007].

ORCID

Wenwu Shen http://orcid.org/0000-0002-8576-4872

References

[1] Liu N, Truong V, Wang X, et al. Integrated scheduling and capacity planning with considerations for patients' length-of-stays. Prod Oper Manag. 2019;28:1735–1756.

[2] Dellaert N, Jeunet J. A variable neighborhood search algorithm for the surgery tactical planning problem. Comput Oper Res. 2017;84:216–225.

[3] Fgener A, Hans EW, Kolisch R, et al. Master surgery scheduling with consideration of multiple downstream units. Eur J Oper Res. 2014;239(1):227–236.

[4] Vanberkel PT, Boucherie RJ, Hans EW, et al. A survey of health care models that encompass multiple departments. Int J Health Manag Inf. 2009;1(29):37–69.

[5] Kumar A, Mo J. Models for bed occupancy management of a hospital in Singapore; 2013. p. 1–6.

[6] Robb WB, O'Sullivan MJ, Brannigan AE, et al. Are elective surgical operations cancelled due to increasing medical admissions? Ir J Med Sci. 2004;173(3):129–132.

[7] Teoh SY, Pan SL, Ramchand AM. Resource management activities in healthcare information systems: a process perspective. Inf Syst Front. 2012;14(3):585–600.

[8] Thomas BG, Bollapragada S, Akbay K, et al. Automated bed assignments in a complex and dynamic hospital environment. Interfaces. 2013;43(5):435–448.

[9] Kim SC, Horowitz I. Scheduling hospital services: the efficacy of elective-surgery quotas. Omega. 2002;30(5):335–346.

[10] Smith D, Vicki L, Schweikhart SB, Dwight E. Capacity management in health care services: review and future research directions. Decis Sci. 1988;19(4):889–919.

[11] Ceschia S, Schaerf A. Local search and lower bounds for the patient admission scheduling problem. Comput Oper Res. 2011;38(10):1452–1463.

[12] Shukla RK. Admissions monitoring and scheduling to improve work flow in hospitals. Inquiry. 1985;22(1):92.

[13] Bowers J. Balancing operating theatre and bed capacity in a cardiothoracic centre. Health Care Manag Sci. 2013;16(3):236–244.

[14] Kolker A. Process modeling of ICU patient flow: effect of daily load leveling of elective surgeries on ICU diversion. J Med Syst. 2009;33(1):27–40.

[15] Min D, Yih Y. Scheduling elective surgery under uncertainty and downstream capacity constraints. Eur J Oper Res. 2010;206(3):642–652.

[16] van Oostrum JM, Van Houdenhoven M, Hurink JL, et al. A master surgical scheduling approach for cyclic scheduling in operating room departments. OR Spectrum. 2008;30(2):355–374.

[17] Vanberkel PT, Boucherie RJ, Hans EW, et al. An exact approach for relating recovering surgical patient workload to the master surgical schedule. J Oper Res Soc. 2011;62(10):1851–1860.

[18] Ma GX, Demeulemeester E. A multilevel integrative approach to hospital case mix and capacity planning. Comput Oper Res. 2013;40(9):2198–2207.

[19] Theresia van Essen J, Bosch JM, Hans EW, et al. Reducing the number of required beds by rearranging the OR-schedule. OR Spectrum. 2014;36(3):585–605.

[20] Samudra M, Van Riet C, Demeulemeester E, et al. Scheduling operating rooms: achievements, challenges and pitfalls. J Sched. 2016;19(5):493–525.

[21] Blake JT. An introduction to platelet inventory and ordering problems. Wiley Encyclopedia of Operations Research and Management ScienceJohn Wiley & Sons; 2011.

[22] Dexter F, Traub RD. How to schedule elective surgical cases into specific operating rooms to maximize the efficiency of use of operating room time. Anesth Analg. 2002;94(4):933.

[23] Mcmanus ML, Long MC, Cooper A, et al. Queuing theory accurately models the need for critical care resources. Anesthesiology. 2004;100(5):1271–1276.

[24] Blake JT, Carter MW. A goal programming approach to strategic resource allocation in acute care hospitals. Eur J Oper Res. 2002;140(3):541–561.

[25] Koizumi N, Kuno E, Smith TE. Modeling patient flows using a queuing network with blocking. Health Care Manag Sci. 2005;8(1):49.

[26] Chow VS, Puterman ML, Salehirad N, et al. Reducing surgical ward congestion through improved surgical scheduling and uncapacitated simulation. Prod Oper Manag. 2011;20(3):418–430.

[27] Zhu YH, Toffolo TAM, Vancroonenburg W. Compatibility of short and long term objectives for dynamic patient admission scheduling. Comput Oper Res. 2019;104:98–112.

[28] Ceschia S, Schaerf A. Dynamic patient admission scheduling with operating room constraints, flexible horizons, and patient delays. J Sched. 2016;19:377–389.

[29] Lusby RM, Schwierz M, Range TM, et al. An adaptive large neighborhood search procedure applied to the dynamic patient admission scheduling problem. Artif Intell Med. 2016;74:21–31.

[30] Rockafellar RT. Convex analysis. 2nd ed. Hoboken: Princeton University; 1972.

[31] Porteus EL. Foundations of stochastic inventory theory. Redwood City: Stanford University Press; 2002.

[32] Huh WT, Liu N, Truong VA. Multiresource allocation scheduling in dynamic environments. Manuf Serv Oper Manag. 2013;15(2):280–291.