Computers & Operations Research 74 (2016) 78–91

Contents lists available at ScienceDirect

Computers & Operations Research

journal homepage: www.elsevier.com/locate/caor

Stochastic programming analysis and solutions to schedule overcrowded operating rooms in China

Guanlian Xiao[a,c], Willem van Jaarsveld[b], Ming Dong[a], Joris van de Klundert[c,*]

[a] Department of Operations Management, Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai, China
[b] School of Industrial Engineering, Eindhoven University of Technology, Eindhoven, Netherlands
[c] Department of Health Services Management and Organization, Institute of Health Policy and Management, Erasmus University Rotterdam, Rotterdam, Netherlands

ARTICLE INFO

Article history:
Received 18 November 2015
Received in revised form
13 April 2016
Accepted 13 April 2016
Available online 19 April 2016

Keywords:
Operating room scheduling
3-Stage stochastic programming
Sample average approximation
China's hospital reform

ABSTRACT

As a result of the growing demand for health services, China's large city hospitals have become markedly overstretched, resulting in delicate and complex operating room scheduling problems. While the operating rooms are struggling to meet demand, they face idle times because of (human) resources being pulled away for other urgent demands, and cancellations for economic and health reasons. In this research we analyze the resulting stochastic operating room scheduling problems, and the improvements attainable by scheduled cancellations to accommodate the large demand while avoiding the negative consequences of excessive overtime work. We present a three-stage recourse model which formalizes the scheduled cancellations and is anticipative to further uncertainty. We develop a solution method for this three-stage model which relies on the sample average approximation and the L-shaped method. The method exploits the structure of optimal solutions to speed up the optimization. Scheduled cancellations can significantly and substantially improve the operating room schedule when the costs of cancellations are close to the costs of overtime work. Moreover, the proposed methods illustrate how the adverse impact of cancellations (by patients) for economic and health reasons can be largely controlled. The (human) resource unavailability however is shown to cause a more than proportional loss of solution value for the surgery scheduling problems occurring in China's large city hospitals, even when applying the proposed solution techniques, and requires different management measures.

1. Introduction

In the first decade of the present millennium, China's GDP has grown at an average rate of more than 10% [35]. These economic developments have gone hand in hand with social and demographic developments. The urban population grew from 452 million to 721 million [45], the public transportation system improved considerably, and health insurance coverage grew from below 30% around the turn of the millennium to over 95% in 2011 [34]. These changes have driven an enormous growth in demand for health services, and in health expenditures of which 71% are accounted for by hospitals [2]. As a result of these developments, particularly the demand for services at the large (level 3) hospitals in big cities increased [38]. Despite a tenfold growth in government spending on health [28] and a growth in the number of hospitals by more than 40% since the year 2000 [36], the increase in health service capacity has not been able to cope with the rising demand. The level 3 hospitals in big cities have become markedly overstretched [41]. These phenomena are concretely illustrated by the 2013 data provided for the purpose of the analysis presented in this manuscript by Shanghai General Hospital, where the actual average surgical workload exceeded the daily capacity by as much as 20%, and average operating room opening hours are almost 14 h daily.

Because a referral system is lacking, an important part of the increased demand directly reaches the hospitals in the form of ever higher numbers of outpatients, which tend to pull away physicians and other staff from wards and operating rooms. The number of outpatient visits to hospitals has grown from 2.12 billion per year to 3.45 billion per year in the first decade of the new millennium [29]. The increase in outpatient services may cause physicians to be late for operating room shifts or to be called away during operating room shifts, causing idle time at the operating room. (From the complete operating room data for the year 2013, we estimate that idle time at Shanghai General Hospital is around 17%.) In the same decade, the number of inpatient visits in China has more than doubled from 53 million to 133 million annually. Meara et al. [33] recently conservatively estimated the annually

[*] Corresponding author.
E-mail addresses: xglian@gmail.com (G. Xiao), w.l.v.jaarsveld@tue.nl (W. van Jaarsveld), mdong@sjtu.edu.cn (M. Dong), vandeklundert@bmg.eur.nl (J. van de Klundert).

http://dx.doi.org/10.1016/j.cor.2016.04.017

needed number of surgeries in China at 57 million, of which they considered 27 million to be unmet. The already overstretched operating rooms are therefore likely to face considerable further increases in demand in the coming years. Hence Chinese hospitals face severe operational problems, now and in the coming years.

Our aim is to develop scheduling methods to solve the urgent capacity management problems in China's large city hospitals – which form a priority in the current health system reform – and see how their effectiveness interacts with accompanying operations management measures. As we outline more extensively in the literature review in the next section, current losses of scarce capacity are mostly not due to poor scheduling, but to other causes such as unavailability of scarce (human) resources, and *cancellations* of planned surgeries. Scheduled surgeries may be canceled for a variety of reasons which are beyond the locus of control of operating room management, such as no-show, deteriorating health conditions, and hospital logistics. In anticipation of such *exogenous cancellations*, operating room management may choose to schedule more patients than capacity allows, potentially resulting in capacity problems when cancellations are fewer than expected, or surgeries take longer than expected. The schedulers may subsequently solve the capacity problems by cancelling one or more of the final patients for which surgery was scheduled at the end of the day. Such cancellations may cause dissatisfaction, anxiety and loss of health for the patients, and have led to tense relations between patients and staff. The alternative to further extend overtime hours, on the other hand, is associated with increased risks of complications and medical errors, as well as dissatisfaction among scarce staff ([39] and references therein). The scheduling of operating rooms in the overstretched Chinese hospitals is therefore a stochastic balancing act which is complicated by resource unavailability and exogenous cancellations.

Operating room schedules are typically constructed one or several days in advance. Because of the stochastic nature of surgical services and the related health service processes, schedules are subsequently often adjusted as the day progresses. For operating rooms for elective surgeries, such adjustments are primarily constrained to changes in surgery start times and, when needed, to cancellations of one or more surgeries of the final patients of the day. It is preferable to take such scheduling decisions to cancel one or more of the final patients early, so as to limit the negative effects for patients and staff mentioned above. In practice, such cancellations may also take the form of redirecting patients to another hospital.

The first research objective is now to optimize the operating room schedules. This starts with the optimization of the schedules created one or more days in advance per single elective operating room, henceforth referred to as the first stage problem. Secondly, we consider the optimization of early *scheduled cancellations*, cancellations initiated by the operating room schedulers after an initial part of the daily schedule has been completed (see for instance [39]), referred to as the second stage problem. In particular, we analyze the improvements attainable by introducing a two stage approach (in which the first stage solution takes into account that a second stage follows) over the common practice of a single stage approach which disregards cancellation until the end of the day. The objective will be to balance the benefits from performing surgeries with the costs of overtime work and negative effects of scheduled cancellations. Our modelling of overtime costs reflects the empirical findings that overtime work is increasingly undesirable for patients and staff as the duration lengthens. Moreover, we model resource unavailability and exogenous cancellations as independent stochastic processes and consider surgical durations to be stochastic as well, fitting real life data. As we are interested in the performance improvement possible by adopting a two stage approach, we develop solution methods which solve the problem with and without scheduled cancellations (almost) to optimality. (See Fig. 1 in Section 3.1 for a visualization of the multistage model.)

With these solution methods at hand, the second research objective is then to analyze the extent to which scheduling can overcome the difficulties posed by stochastic resource unavailabilities and exogenous cancellations or, alternatively, whether additional operations management measures are required for this purpose. This second research objective is particularly relevant as the literature review below shows that resource unavailability and exogenous cancellations are, to a certain extent, under the control of hospital management. Hence, our results provide insight in how operating room scheduling and hospital management can interact to alleviate China's hospital overcrowding problems.

Section 2 reviews related literature on (surgical) scheduling with cancellations as well as literature on the occurrence and causes of surgical cancellation. Section 3 formally defines the problem and formulates it as a general three-stage model with integer recourse. Section 4 analyzes theoretical model properties which can help to reduce solution times. Section 5 proposes specific solution algorithms for the problem, and finally Section 6 presents numerical results and analysis. The numerical analysis tests the newly developed 3-stage stochastic programming approach by (almost) optimally solving instances derived from 2013 operating room data of Shanghai General Hospital. To this purpose, we fit distributions to the underlying stochastic processes using a complete data set on surgical operations. QQ-plots show that lognormal distributions fit these surgical durations well, and the proposed SAA approach is able to deal with these analytically inconvenient distributions. The computational results provide insight in the benefits attainable by scheduled cancellations for current rates of resource unavailability and exogenous cancellations. Moreover, we consider scenarios in which additional measures are taken to reduce resource unavailability and exogenous cancellations. We conclude by considering practical implications for operating room management and scheduling in China's overcrowded hospitals.

2. Literature review

The phenomena of cancellation, no-show and overbooking have been studied extensively in the operations management literature, mostly originating from revenue management applications in the airline industry [42]. In this setting, no-show refers to passengers not showing up for a flight without giving prior notice, and cancellation to passengers cancelling their booked flights in advance (which is different from the definitions for cancellations provided above). Like it is the case in the surgical scheduling problem we consider, revenue management models typically exploit the expected benefits from overbooking capacity, taking into account that penalties must be paid when the eventual number of patients showing up exceeds capacity. For instance Subramanian et al. [40] consider an application which includes no-show, cancellation and overbooking. While the revenue management problems considered in the airline and hotel industry are essentially different from surgical scheduling, they share general properties and solution approaches. For instance, Karaesmen and Van Ryzin [20] present a two-stage stochastic program to model no-show and overbooking, where cancellations have become known in the second stage (as is partially the case in our model). Lai and Ng [25] propose a stochastic network optimization model for hotel revenue management and use robust optimization techniques to deal with cancellations, no-show and over-booking of hotel guests. Overbooking has also been introduced in health care, first and foremost in appointment scheduling for outpatients. For instance LaGanga and Lawrence [24] and Berg et al. [4] use overbooking to hedge against patient no-show and present simulation results showing a significant improvement in access and provider productivity, while increasing both patient wait times and provider overtime.

With regard to surgery scheduling, May et al. [32] conclude from a literature review that 'it remains to be seen if the existing results and observations regarding manufacturing replanning and rescheduling would extend to surgery' (where rescheduling refers to the possibility to adjust the initial schedule during execution). Much of the literature on surgical scheduling optimizes the sequence and schedule for a fixed pool of patients while taking the stochastic nature of several problem parameters, especially surgery duration, into account. Mancilla and Storer [31], Denton et al. [12] and Berg et al. [4] simultaneously consider patient waiting time, resource idle time, and overtime. Xiao et al. [47] propose an adaptive scheduling approach for a problem that is closely related to the one considered in this paper, yet without considering cancellation. Stepaniak et al. [39] present a simulation study on cancellation, which they refer to as 'patient rejection'. Formal scheduling models which explicitly include cancellation, as is particularly relevant for overcrowded hospitals, appear to have received little or no attention in operating room scheduling so far.

The scheduling process we adopt matches a multiple stage stochastic programming approach. Standard two-stage stochastic programs with linear or convex functions are often solved using the L-shaped method or Bender's decomposition [44,6,7]. However, our recourse decision (scheduled cancellations) is still anticipative to further uncertainty, namely the second shift surgery durations, unavailability and cancellations. As such, the decision problem can be viewed as a three-stage recourse model [5,6]. Solving the scheduling problem is further complicated because the recourse function is integer. Laporte and Louveaux [26] propose modified L-shaped decomposition with adjusted optimal cuts for two stage stochastic program with integer recourse. Angulo et al. [1] alternately generate optimal cuts of the linear sub-problem and the integer sub-problem, which improves the practical convergence (see also [15,8]). We follow a sample average approximation approach (SAA) which uses this framework. Moreover, we prove and exploit a specific relationship between the first-stage realization and the optimal number of scheduled cancellations to speed up the computation of integer cuts. We use Jensen's inequality [17] to upper bound the minus second (and third) stage cost, a technique that was proposed by Batun et al. [3].

We now review studies on the occurrence and cause of surgery cancellations. Cancellation of surgery is a common phenomenon globally and appears to be more frequent in developing counties. For instance, Kumar and Gandhi [23] (India) report that 17.6% of scheduled surgeries are canceled on the day of surgery. Several authors, e.g., Kumar and Gandhi [23], Kolawole and Bolaji [22] (Nigeria), Chiu et al. [10] (China), Chalya et al. [9] (Tanzania), analyze causes of cancellation, citing variations and prolonged durations of previous surgeries as a prime source. A Daily Briefing [11] report discusses a case study in the USA in which 6.7% of scheduled surgeries in 2009 are canceled, one-third of which was due to hospital related causes, such as poor scheduling. In addition, Yoon et al. [49] (Korea), Hussain and Khan [16] (Pakistan), Perroca et al. [37] (Brazil) and Fernando et al. [14] (UK) explore cancellations. The latter authors point at the management role to address the inefficiencies that cancellations may cause. The Lancet Commission on Global Surgery posits that management might be even more important in settings in which maximal use of the few available resources is a practical necessity to advance on meeting the unmet global need of 143 million surgeries yearly [33].

Various authors report cancellation rates of between 10% and 15% for Chinese hospitals. Jiang et al. [18] report that 12.88% of children's elective surgeries are canceled in Hunan children's hospital in 2010 due to emergent infection (70.30%), inappropriate preoperative preparation (15.12%), poor scheduling and other factors (14.58%). Jie et al. [19] take a statistical analysis on Guangdong General Hospital, which is a large general hospital, and show that the cancellation rate is at 11.2%. Causes for cancellations are patients' illnesses (65.97%), lack of preoperative preparations (14.03%), economic reasons and risk concerns (10.99%), and accidents (9.01%). (Economic reasons refer to the patients inability to pay.) Li et al. [27] study cancellation at Zunyi Medical College, and report as main causes of cancellation: upper respiratory tract infection (18.39%), high blood pressure (12.86%), lack of preoperative preparation (11.79%), and economic concerns (9.64%). Xiang et al. [46] report a cancellation rate of 5.1% caused by recent changes in health conditions (55.8%), patients' determination changes (23.1%), and poor scheduling. Zhang et al. [50] report a 2010 case study and find that the cancellation rate is 13.9%, due to illnesses (68.7%), exogenous cancellations (20.3%), and preoperative preparations (7.7%). The reader may refer to Xu et al. [48] for related work. Next to scheduling related reasons, several of these authors mention the length of schedules and workload as reasons for *scheduled cancellations*.

Briefly reflecting on these causes of cancellations, we notice that they are mostly attributed to emergent infection, illness, recent changes in health condition and the like. It is not uncommon that these conditions relate to hospital acquired infections, which are preventable. Procedures for hospitalization and infection prevention may reduce the prevalence of these cancellations. Another important source of cancellation stems from the high out-of-pocket (co-)payments patients have difficulty to effectuate. Improvements in health insurance coverage, as currently in progression, may reduce the number of these economically driven cancellations. In our computational experiments we explore scenarios in which exogenous cancellations are less frequent.

3. The model

3.1. Problem description and notation

For the single operating room scheduling problem under consideration, we denote by $\hat{t}$ the regular working time. For example, in Shanghai General Hospital, $\hat{t}$ equals 570 min (9.5 h). An initial schedule is made at least one day ahead. This initial schedule specifies a sequence for the patients and expected starting times of their surgeries. The patients to be scheduled are selected from a given set $I = \{1, 2, \dots, n_p\}$. The reward of performing surgery on patient $i \in I$ equals r_i. This reward can be interpreted strictly financially, in which case it corresponds to the associated hospital revenue [13], or can be defined more broadly to incorporate for instance also the benefits for the patients (see also [47]). Notice that in the latter case, the corresponding values may not be readily available from hospital information systems. Scheduled cancellation of surgery for patient $i \in I$ leads to a penalty of c_i, which can in turn be a financial penalty incurred by the insurer, including wasted pre-operative costs, and more generally including patient inconveniences and losses of health.

Each patient $i \in I$ has an associated surgical time distribution, which will be denoted by ξ_i'. We assume that the surgery times for different patients are independent. We also include a probability of exogenous cancellation, which will be denoted by $_{pi}$ for all $i \in I$. There are no rewards for exogenously canceled surgeries and they do not take time except for a constant t_d switching time. For each patient $i \in I$, selecting patient i thus consumes ξ_i' time units of operating theater capacity with probability $1-p_i$, and t_d time units of capacity with probability $_{pi}$. We let ξ_i represent this compound random variable which equals ξ_i' with probability $1-p_i$ and t_d with probability $_{pi}$.

In many practical contexts, a number of patients may have the same characteristics (from the perspective of scheduling), because they have to undergo the same procedure. To accommodate this, if for patients i and i' we have that $r_i = r_{i'}$, $c_i = c_{i'}$, $p_i = p_{i'}$, and that ξ_i' and $\xi_{i'}'$ are identically distributed, then we will say that patients i and i' belong to the same surgery class. This will be denoted by $i \sim i'$. Let it be noted however that ξ_i' and $\xi_{i'}'$ will still be independent. More explicitly, while patients may share characteristics, the surgery time distributions ξ_i' and $\xi_{i'}'$ of each pair of patients i and i' are independent, even if ξ_i' and $\xi_{i'}'$ are identically distributed.

As outlined in the introduction, unavailability of surgical resources (staff and/or facilities) is another important source of uncertainty which reduces the effective time available for surgery in the operating room. We thus introduce η_1 and η_2, which represent the total length of such interruptions in the first and second shift, respectively.

In practice, decision making regarding scheduled cancellations may for instance take place daily at a fixed moment in time (see e.g. [39] for example set this moment at 2 PM). We adopt a different approach, which guarantees a first shift of patients that their surgeries will be scheduled, and allows to inform a second shift of scheduled patients that they will either receive final confirmation or notification of cancellation after the first shift is completed. We consider this approach to be more patient centered as it eliminates uncertainty for the first shift patients and provides clarity to all others after this first shift has been completed. To this purpose, we set the moment of decision making on scheduled cancellations upon completion of half of the scheduled patients (rounded down in case of an odd number of patients). The time of completion of the first shift therefore forms the recourse moment in the proposed multi-stage stochastic programming approach. The second stage thus entails to decide on possible scheduled cancellations of surgeries for patients scheduled in the second shift. After this recourse moment, the second shift surgery durations are revealed and final costs are incurred, making the problem a three-stage recourse model [6].

Following current practice, we assume that scheduled cancellations always regard the last patients in the sequence implied by the surgical schedule, working backwards through the sequence if more than one scheduled patient is canceled. To model the scheduled cancellations we introduce *positions*. All patients scheduled in the first shift are considered to be in position $j=0$, because their order is inconsequential from the viewpoint of our model. For the second shift, we introduce positions $j \in \{1, \ldots, \kappa\}$, that are to be filled sequentially, starting from position 1. We later comment on how to set κ. The set of all positions will be denoted by $\{0, 1, \ldots, \kappa\}$; this includes the first and second shifts.

We introduce binary decision variables x_{ij}, $i \in I, j \in \{0, 1, \ldots, \kappa\}$, where x_{ij} equals 1 if patient i is scheduled in the jth position, and 0 otherwise. For convenience, let $\mathbf{x} = \{x_{ij} | i \in I, j \in \{0, 1, \ldots, \kappa\}\}$. By interpretation, $\sum_{i=1}^{n_p} x_{i0}$ represents the number of patients scheduled for the first shift. Second shift slots $j \in \{1, \ldots, \kappa\}$ may contain at most a single patient. To balance the patient numbers between the shifts as described above, we use the restriction $\sum_{i=1}^{n_p} x_{i0} = \lfloor \sum_{i=1}^{n_p} \sum_{j=0}^{\kappa} x_{ij}/2 \rfloor$, where $\lfloor x \rfloor$ is the largest integer no greater than x. We thus need no more than $\lfloor n_p/2 \rfloor + 1$ second shift positions, and may set $\kappa = \lfloor n_p/2 \rfloor + 1$ accordingly.

To specify the three-stage recourse model with (integer) recourse, we create i.i.d. copies s_i of each random variable ξ_i, which will represent the surgery times in the first shift. Variables s_i and ξ_i follow the same distribution but are independent. We then denote the first shift of the schedule by $\mathbf{s} = (s_1, s_2, \ldots, s_{n_p}, \eta_1)$, and the second shift by $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{n_p}, \eta_2)$. We set the rewards for patients corresponding to exogenous cancellations to zero. Thus, reward loss due to exogenous cancellations can be modeled as $\sum_{i=1}^{n_p} I_0(s_i) r_i x_{i0}$, and the indicator function $I_0(s_i) = 1$ if $s_i = t_d$, and 0 otherwise. Next consider the binary decision variables $y_{ij}(\mathbf{s})$, $i \in I, j \in \{1, \ldots, \kappa\}$, which depend on the outcome of $\mathbf{s}$. We let $y_{ij}(\mathbf{s}) = 1$ if treatment of patient i in slot j is canceled under scenario $\mathbf{s}$, and $y_{ij}(\mathbf{s}) = 0$ otherwise. For convenience, let $\mathbf{y} = \{y_{ij} | i \in I, j \in \{1, \ldots, \kappa\}\}$. Scheduled cancellation of patient i results in a penalty r_i'. Moreover, scheduled cancellations require zero time. The total amount of time that schedule $(\mathbf{x}, \mathbf{y})$ takes is therefore $\sum_{i=1}^{n_p} s_i x_{i0} + \eta_1 + \sum_{i=1}^{n_p} \sum_{j=1}^{\kappa} [x_{ij} - y_{ij}(\mathbf{s})]\xi_i + \eta_2$. The loss of reward in the second stage due to exogenous cancellation is $\sum_{i=1}^{n_p} \sum_{j=1}^{\kappa} [x_{ij} - y_{ij}(\mathbf{s})] r_i I_0(\xi_i)$.

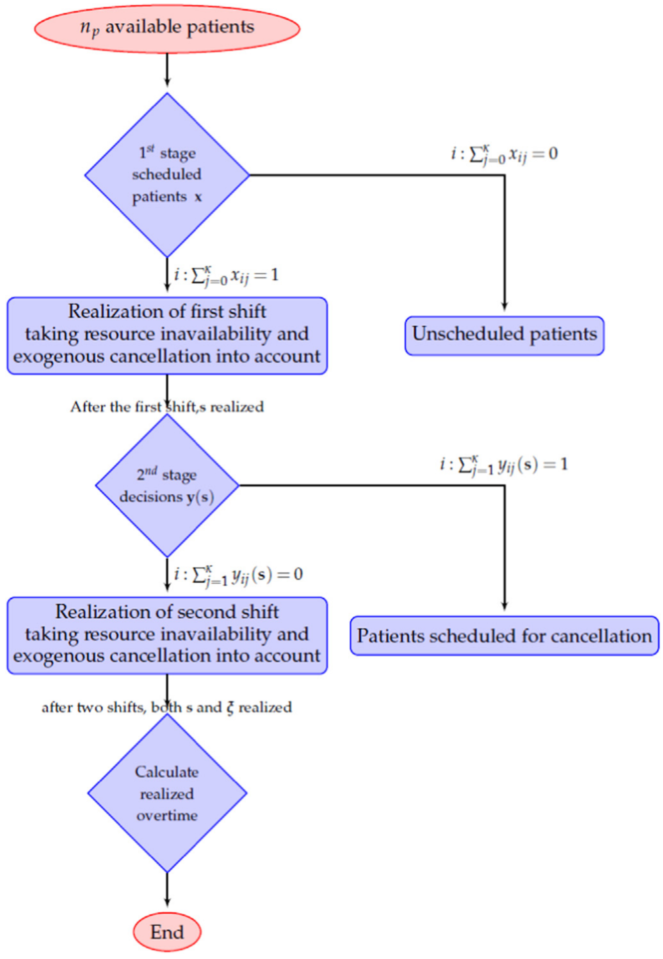
n_p available patients
1^{st} stage scheduled patients x
$i: \sum_{j=0}^{\kappa} x_{ij} = 0$
$i: \sum_{j=0}^{\kappa} x_{ij} = 1$
Realization of first shift taking resource inavailability and exogenous cancellation into account
Unscheduled patients
After the first shift,s realized
2^{nd} stage decisions y(s)
$i: \sum_{j=1}^{\kappa} y_{ij}(s) = 1$
$i: \sum_{j=1}^{\kappa} y_{ij}(s) = 0$
Realization of second shift taking resource inavailability and exogenous cancellation into account
Patients scheduled for cancellation
after two shifts, both s and ξ realized
Calculate realized overtime
End

Fig. 1. A chart showing the flow of the patients in the various decision stages in our scheduling problem.

We assume that overtime work incurs a cost, which may include financial costs such as salary, employee dissatisfaction, and patient safety risks, which increase with the duration of overtime (see also Section 1). We therefore model the overtime cost function to be piecewise linear and convex, as illustrated in the example in Fig. 2. In the example overtime starts after 570 min and overtime cost per time unit becomes more expensive per time unit after 120 min of overtime.

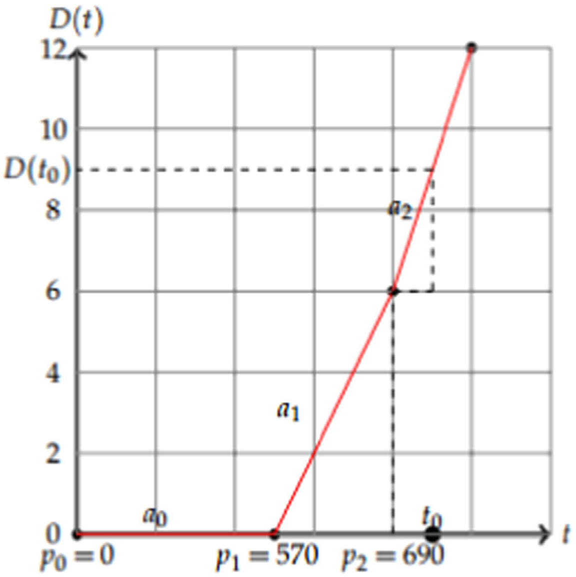

Fig. 2. Piecewise linear and convex function $D(t)$. Penalty cost $D(t)$ as a function of working time t as illustrated for (t_0).

3.2. Stochastic programming formulation

We now formulate the scheduling problem as a stochastic program with recourse. For ease of reference, we repeat that $\mathbf{s} = (s_1, s_2, \dots, s_{n_p}, \eta_1)$ and $\boldsymbol{\xi} = (\xi_1, \dots, \xi_{n_p}, \eta_2)$ are the random variables pertaining to the first and second shift, respectively. Note that the recourse decision (scheduled cancellations) must be made after $\mathbf{s}$ is revealed, but based on distributional information on ξ alone. We obtain the following formulation:

$$\max_{\mathbf{x}\in X} \sum_{i=1}^{n_p} \sum_{j=0}^{\kappa} r_i \cdot x_{ij} - E_{\mathbf{s}} Q(\mathbf{x}, \mathbf{s}) \tag{1}$$

where

$$\begin{aligned} Q(\mathbf{x}, \mathbf{s}) = & \sum_{i=1}^{n_p} I_0(s_i) r_i x_{i0} + \min_{\mathbf{y}(\mathbf{s})\in Y(\mathbf{x})} \sum_{i=1}^{n_p} \sum_{j=1}^{\kappa} c_i y_{ij}(\mathbf{s}) \\ & + E_{\xi}[D\left(\sum_{i=1}^{n_p} s_i x_{i0} + \eta_1 + \sum_{i=1}^{n_p} \sum_{j=1}^{\kappa} [x_{ij} - y_{ij}(\mathbf{s})]\xi_i + \eta_2 \right) \\ & + \sum_{i=1}^{n_p} \sum_{j=1}^{\kappa} I_0(\xi_i) r_i [x_{ij} - y_{ij}(\mathbf{s})]]. \end{aligned} \tag{2}$$

X and $Y(\mathbf{x})$ will be detailed below: They represent the feasible domain for the first and second stage decisions, respectively. In particular, we have

$X = \{\mathbf{x} | (3)–(8)\}$

$$\sum_{i=1}^{n_p} x_{i1} \le 1 \tag{3}$$

$$\sum_{i=1}^{n_p} x_{ij+1} - \sum_{i=1}^{n_p} x_{ij} \le 0, \quad \forall\, j \in \{1, \dots, \kappa - 1\} \tag{4}$$

$$0 \le \sum_{i=1}^{n_p} \sum_{j=1}^{\kappa} x_{ij} - \sum_{i=1}^{n_p} x_{i0} \le 1 \tag{5}$$

$$\sum_{j=0}^{\kappa} x_{ij} \le 1, \quad \forall\, i \in I \tag{6}$$

$$x_{ij+1} \le \sum_{k=1}^{j} x_{i'k}, \quad \forall\, i, i' \in I: i \sim i', i < i', \;\; j \in \{0, \dots, \kappa - 1\} \tag{7}$$

$$x_{ij} \in \{0, 1\}, \quad \forall\, i \in I, \;\; j \in \{0, \dots, \kappa\} \tag{8}$$

Combining (3) and (4) ensures that second shift positions $j \in \{1, \dots, \kappa\}$ are filled sequentially, and with at most a single patient. The workload is balanced by (5), which ensures that the number of patients scheduled in the first shift is equal to the number of patients in the second shift, or one less. Each patient is scheduled at most once by (6). While (7) is not necessary, it greatly reduces the search space by reducing symmetry.

The feasible domain for the second stage decisions depends on the first stage decision $\mathbf{x}$, and is given by:

$Y(\mathbf{x}) = \{\mathbf{y} | (9)–(11)\}$

$$y_{i'j} - \sum_{i=1}^{n_p} y_{ij+1} \le 1 - \sum_{i=1}^{n_p} x_{ij+1}, \quad \forall\, i' \in I, \quad \forall\, j \in \{1, \dots, \kappa - 1\} \tag{9}$$

$$y_{ij} \le x_{ij}, \quad \forall\, i \in I, \;\; j \in \{1, \dots, \kappa\} \tag{10}$$

$$y_{ij} \in \{0, 1\}, \quad \forall\, i \in I, \;\; j \in \{1, \dots, \kappa\} \tag{11}$$

We may not cancel a patient in a position unless all patients with higher position are also canceled, which is enforced by (9). Indeed, if a patient is scheduled in position $j + 1$, then $\sum_{i=1}^{n_p} x_{ij+1} = 1$, and (9) enforces that a treatment at position j can only be canceled if a treatment at position $j + 1$ is canceled as well. If no patient is scheduled at position $j + 1$, then $\sum_{i=1}^{n_p} x_{ij+1} = 0$, and we are free to cancel the treatment at position j. Only patients who are actually scheduled may be canceled, which is enforced by (10).

For later convenience, define $\bar{X}$ and $\bar{Y}$ as the continuous relaxation of X and Y, respectively. Hence, $\bar{X} = \{\mathbf{x} \big| ((3)–(7)) + (12)\}$, with

$$x_{ij} \in [0, 1], \quad \forall\, i \in I, \;\; j \in \{0, \dots, \kappa\} \tag{12}$$

and $\bar{Y}(\mathbf{x}) = \{\mathbf{y} \big| ((9)–(10)) + (13)\}$, with

$$y_{ij} \in [0, 1], \quad \forall\, i \in I, \;\; j \in \{1, \dots, \kappa\} \tag{13}$$

3.3. A different formulation of the second-stage problem

For any first stage solution $\mathbf{x}$, let $k_{\max} = \sum_{i=1}^{n_p} \sum_{j=1}^{\kappa} x_{ij}$. We now present an equivalent formulation of the second stage problem $Q(\mathbf{x}, \mathbf{s})$:

$$\begin{aligned} \tilde{Q}(\mathbf{x}, \mathbf{s}) = & \sum_{i=1}^{n_p} I_0(s_i) r_i x_{i0} + \min_{k_{\mathbf{s}} \in Z, 0 \le k_{\mathbf{s}} \le k_{\max}} \sum_{i=1}^{n_p} \sum_{j=k_{\mathbf{s}}+1}^{\kappa} c_i x_{ij} \\ & + E_{\xi}[D\left[\sum_{i=1}^{n_p} s_i x_{i0} + \eta_1 + \sum_{i=1}^{n_p} \sum_{j=1}^{k_{\mathbf{s}}} \xi_i x_{ij} + \eta_2 \right] \\ & + \sum_{i=1}^{n_p} \sum_{j=1}^{k_{\mathbf{s}}} I_0(\xi_i) r_i x_{ij}], \end{aligned} \tag{14}$$

(where $\sum_{j=1}^{k_{\mathbf{s}}} := 0$ when $k_{\mathbf{s}} = 0$). Clearly, the decision variable $k_{\mathbf{s}}$, which appears as a summation index, makes this formulation non-standard and less suitable for computational purposes. The formulation nevertheless allows to deduce some structural properties.

Lemma 1. *The second stage decision problems* $Q(\mathbf{x}, \mathbf{s})$ *and* $\tilde{Q}(\mathbf{x}, \mathbf{s})$ *are equivalent for any* $\mathbf{x} \in X$ *and realization of* $\mathbf{s}$.

All proofs of lemmas and propositions are provided in Appendix A. This lemma yields the following equivalent formulation of (1), which will be analyzed in the next section:

$$\max_{\mathbf{x} \in X} \sum_{i=1}^{n_p} \sum_{j=0}^{\kappa} r_i \cdot x_{ij} - E_{\mathbf{s}} \tilde{Q}(\mathbf{x}, \mathbf{s}) \tag{15}$$

4. Analytical Insights

4.1. Structural properties of the second stage problem

In this section we develop a relation between the capacity used by the first shift and the cancellations in the second shift for a fixed schedule $\mathbf{x} \in X$. Firstly, we introduce some notations:

$$\hat{s} = \sum_{i=1}^{n_p} x_{i0} s_i + \eta_1 \tag{16}$$

$$R(\mathbf{s}) = \sum_{i=1}^{n_p} I_0(s_i) r_i x_{i0} \tag{17}$$

$$g(\hat{s}, k_{\hat{s}}) = \sum_{i=1}^{n_p} \sum_{j=k_{\hat{s}}+1}^{\kappa} c_i x_{ij} + E_{\xi}\left[D\left[\hat{s} + \sum_{i=1}^{n_p} \sum_{j=1}^{k_{\hat{s}}} \xi_i x_{ij} + \eta_2 \right] + \sum_{i=1}^{n_p} \sum_{j=1}^{k_{\hat{s}}} I_0(\xi_i) r_i x_{ij} \right] \tag{18}$$

$$f(\hat{s}) = \min_{0 \le k_{\hat{s}} \le k_{\max}} g(\hat{s}, k_{\hat{s}}) \tag{19}$$

$$k_{\hat{s}}^* = \max[\arg \min_{0 \le k_{\hat{s}} \le k_{\max}} g(\hat{s}, k_{\hat{s}})] \tag{20}$$

Thus, we let $\hat{s}$ denote the total realized time of the first shift for a given first stage solution $\mathbf{x}$ and we let $g(\hat{s}, k_{\hat{s}})$ be the corresponding second stage cost (excluding $R(\mathbf{s})$) when $k_{\hat{s}}$ patients are kept in the second shift. By $k_{\hat{s}}^*$, we denote the optimal number of patients to keep (not scheduled for cancellation), i.e., the index minimizing $g(\hat{s}, k_{\hat{s}})$, choosing the largest possible index in case of a tie. The associated minimum cost is denoted by $f(\hat{s})$.

Proposition 1. *Let* $\mathbf{x} \in X$ *be given, and conditioned on* $\hat{s}$, *then* $g(\hat{s}, k_{\hat{s}})$ *is a supermodular function.*

With Proposition 1 at hand, we can then prove that:

Proposition 2. *Let* $\mathbf{x} \in X$ *be given, and consider two realizations of the total time of the first shift:* $\hat{s}_1$ *and* $\hat{s}_2$ *with* $\hat{s}_1 \le \hat{s}_2$, *then* $k_{\hat{s}_1}^* \ge k_{\hat{s}_2}^*$.

Since $Q(\mathbf{x}, \mathbf{s})$ and $\tilde{Q}(\mathbf{x}, \mathbf{s})$ are equivalent by Lemma 1, the intuitive practical interpretation of this result is that the number of scheduled cancellations increases with the length of realization of the first shift. The result will be used in the L-shaped method to accelerate the solution of the integer subproblem.

We now rewrite the second stage cost function (18) conditioned on $\hat{s}$ as follows: $F(\hat{s}, \mathbf{y}) = \sum_{i=1}^{n_p} \sum_{j=1}^{\kappa} c_i \cdot y_{ij} + E_{\xi}[D[\hat{s} + \sum_{i=1}^{n_p} \sum_{j=1}^{\kappa} \xi_i (x_{ij} - y_{ij}) + \eta_2] + \sum_{i=1}^{n_p} \sum_{j=1}^{\kappa} I_0(\xi_i) r_i (x_{ij} - y_{ij})]$. Because the L-shaped method requires convexity, the following result is helpful to solve the relaxed model with continuous recourse:

Lemma 2. *Let* $\mathbf{x} \in \bar{X}$ *be given and* $\hat{s}$ *be defined by* (16), *then* $F(\hat{s}, \mathbf{y})$ *is convex in* $\mathbf{y} \in \bar{Y}(\mathbf{x})$.

Observing that $R(\mathbf{s})$ is independent of $\mathbf{y}$, we therefore also have that the second stage objective function is convex in $\mathbf{y} \in \bar{Y}(\mathbf{x})$. The convexity of the second stage objective function in $\mathbf{y} \in \bar{Y}(\mathbf{x})$ will be used in the L-shaped method in Section 5.3 to approximately evaluate the original subproblem with integer recourse.

We conclude this section by a general convexity result for the minimum cost function of the continuous relaxation of the second stage problem, which is further used in Section 4.2.

Proposition 3. *Let* $\mathbf{x} \in X$ *be given and* $\hat{s}$ *be defined by* (16), *then* $f(\hat{s}) = \min_{\mathbf{y} \in \bar{Y}} F(\hat{s}, \mathbf{y})$ *and* $f(\hat{s})$ *is convex in* $\hat{s}$. *Besides,* $\sum_{i=1}^{n_p} I_0(s_i) r_i x_{i0}$ *is also convex in* $\mathbf{s}$.

4.2. Convexity of the second stage problem

We now proceed to derive optimality cuts for the integral master problem and its continuous relaxation on the basis of Jensen's inequality. By Proposition 3 and Lemma 1, we can apply Jensen's inequality [17] to obtain

$$E(f(\hat{s}) + R(\mathbf{s})) \ge f(E(\hat{s})) + R(E(\mathbf{s})) \tag{21}$$

By definition, $\min_{\mathbf{y} \in Y} F(\hat{s}, \mathbf{y}) \ge f(\hat{s})$ for $\forall\, \hat{s}$. Now, by taking expectation on both sides and using inequality (21), we can further derive that

$$E(\min_{\mathbf{y} \in Y} F(\hat{s}, \mathbf{y}) + R(\mathbf{s})) \ge f(E(\hat{s})) + R(E(\mathbf{s})) \tag{22}$$

We will use inequalities (21) and (22) to strengthen our L-shaped algorithm by formulating valid inequalities for continuous and integral master problems, cf. Batun et al. [3].

5. Solution methods

As our research questions require to compare the optimal solutions of various models and parameter settings, we now set out to describe solution techniques designed to present near to optimal solutions. More specifically we present a solution method based on SAA in Section 5.2. Because of the stochasticity still involved after the second stage, we require many samples to accurately represent the stochastic nature of the problem, which makes the SAA approach non-standard and computationally challenging. We use the theoretical results derived in Section 4 to reduce the computation times required to solve the SAA in Section 5.3. In Appendix B.2, the resulting formulation is strengthened using Jensen's inequalities.

5.1. Linearizing the objective function

In order to formulate the SAA as a MIP, we linearize the objective function by writing the overtime cost function as follows:

$$D(x) = \min_{\phi_v} \sum_{v=0}^{q} \tau_v \phi_v$$

$$\text{s.t.} \quad \sum_{v=0}^{q} \phi_v = x$$

$$\phi_v \in [0, l_v]$$

Note that each piecewise linear convex function on $[0, \infty)$ with $q + 1$ breakpoints can be written in this fashion. Here, the length of interval $v \in \{0, \dots, q\}$ is l_v, and its slope is τ_v. The slopes should satisfy $\tau_u \ge \tau_v$ for $u \ge v$.

5.2. SAA formulation

For the SAA, we use $\hat{n}$ independent samples of $\mathbf{s}$, for which we will use the index $n \in \{1, \dots, \hat{n}\}$, and $\hat{m}$ independent samples of $\boldsymbol{\xi}$, for which we will use the index $m \in \{1, \dots, \hat{m}\}$. Denote the first shift surgery time for patient i for sample n by s_{in}, and the time lost due to resource unavailability by η_{1n}. Denote the second shift surgery time and time lost due to resource unavailability for sample m by ξ_{im} and η_{2m}, respectively. Solving the problem consists in finding first stage decisions $\mathbf{x} = \{x_{ij} | i \in I, j \in \{0, 1, \dots, \kappa\}\}$, and for each sample $n \in \{1, \dots, \hat{n}\}$ a second stage decision $\mathbf{y}(n) = \{y_{ij}(n) | i \in I, j \in \{0, 1, \dots, \kappa\}\}$, such that each $\mathbf{y}(n) \in Y(\mathbf{x})$. Here, $\mathbf{y}(n)$ is short for $\mathbf{y}(\mathbf{s}_n)$.

We now formulate the associated sample average approximation (SAA) for (1):

$$\max_{\mathbf{x}\in X} \sum_{i=1}^{n_p} \sum_{j=0}^{\kappa} r_i \cdot x_{ij} - \frac{1}{\hat{n}} \sum_{n=1}^{\hat{n}} Q(\mathbf{x}, n) \tag{23}$$

where

$$Q(\mathbf{x}, n) = \min_{\mathbf{y}(n),l} \quad l + \sum_{i=1}^{n_p} \sum_{j=1}^{\kappa} c_i y_{ij}(n) + \frac{1}{\hat{m}} \sum_{m=1}^{\hat{m}} \left[\left(\sum_{v=0}^{q} \tau_v \phi_v(n, m) \right) - \sum_{i=1}^{n_p} \sum_{j=1}^{\kappa} I_0(\xi_{mi}) r_i y_{ij}(n) \right] \tag{24}$$

$$\text{s.t.} \quad l \ge \sum_{i=1}^{n_p} I_0(s_{in}) r_i x_{i0} + \frac{1}{\hat{m}} \sum_{m=1}^{\hat{m}} \sum_{i=1}^{n_p} \sum_{j=1}^{\kappa} I_0(\xi_{mi}) r_i x_{ij} \tag{25}$$

$$\sum_{v=0}^{q} \varphi_v(n, m) = \sum_{i=1}^{n_p} \sum_{j=1}^{\kappa} \xi_{mi}(x_{ij} - y_{ij}(n)) + \eta_{2m} + \sum_{i=1}^{p} s_{in} x_{i0} + \eta_{1n}, \quad \forall\, n \in \{1, \dots, \hat{n}\}, \quad m \in \{1, \dots, \hat{m}\} \tag{26}$$

$$\varphi_v(n, m) \in [0, l_v], \quad \forall\, n \in \{1, \dots, \hat{n}\}, \quad m \in \{1, \dots, \hat{m}\}, \quad v \in \{0, \dots, q\} \tag{27}$$

$$\mathbf{y}(n) \in Y(\mathbf{x}), \quad \forall\, n \in \{1, \dots, n\} \tag{28}$$

where l is introduced to simplify the formulation. Note that by definition, the set inclusions $\mathbf{x} \in X$ and $\mathbf{y}(n) \in Y(\mathbf{x})$ can be expressed using linear inequalities and binary variables. For example, $\mathbf{y}(n) \in Y(\mathbf{x})$ can be expressed using (9)–(11), with $y_{ij}(n)$ taking the place of y_{ij}. We let $Q(\mathbf{x}) = \frac{\sum_{n=1}^{\hat{n}} Q(\mathbf{x}, n)}{\hat{n}}$.

For the L-shaped method introduced in Section 5.3 we will also use the continuous recourse relaxation $Q_{\text{LP}}(\mathbf{x}, n)$ of $Q(\mathbf{x}, n)$, which is obtained by relaxing (28) to

$$\mathbf{y}(n) \in \bar{Y}(\mathbf{x}). \tag{29}$$

We let $Q_{\text{LP}}(\mathbf{x}) = \frac{\sum_{n=1}^{\hat{n}} Q_{\text{LP}}(\mathbf{x}, n)}{\hat{n}}$.

5.3. Application of L-shaped method

The L-shaped method iteratively generates feasibility and optimality cuts. For the problem under consideration, only optimality cuts are needed. Denote the set of generated optimality cuts by Θ. Each optimality cut provides a lower bound to the second stage cost. That is, for every $\mathbf{x} \in X$ and $(\mathbf{v}_k, \rho_k) \in \Theta$ we have that $Q(\mathbf{x}) \ge \mathbf{v}_k^T\mathbf{x} + \rho_k$ and $-Q(\mathbf{x}) \le -\mathbf{v}_k^T\mathbf{x} - \rho_k$ [26] (here $\mathbf{v}_k^T$ is the transpose of $\mathbf{v}_k$).

$$\max_{\mathbf{x}} \quad \sum_{i=1}^{n_p} \sum_{j=0}^{\kappa} r_i x_{ij} + \theta \tag{30}$$

$$\text{s.t.} \quad \theta \le -\mathbf{v}_k^T\mathbf{x} - \rho_k, \quad \forall\, (\mathbf{v}_k, \rho_k) \in \Theta \tag{31}$$

$$\mathbf{x} \in X \tag{32}$$

Notice that θ bounds the minus of the second stage cost, i.e., θ bounds $-Q(\mathbf{x})$. We will also refer to the *relaxed master problem*, in which (32) is replaced by $\mathbf{x} \in \bar{X}$. In order to strengthen both the master problem and the relaxed master problem, Jensen's inequality is added in the form of an additional constraint involving θ (cf. Appendix B.2).

In the course of our algorithm, we will generate two types of cuts. For the first type, which will be referred to as continuous recourse optimality cuts, we note that for every $\mathbf{x}_l \in \bar{X}$, we can use Benders' decomposition [6] to obtain a cut $(\mathbf{v}, \rho)$ such that $\mathbf{v}^T\mathbf{x}_l + \rho = Q_{\text{LP}}(\mathbf{x}_l)$. That is, at $\mathbf{x}_l$ the cut is tight for the continuous recourse relaxation.

For the second type, which will be referred to as integer optimality cuts, note that for every $\mathbf{x}_l \in X$ we may compute $Q(\mathbf{x}_l)$ by solving the integer second-stage problems. We can then generate a cut $(\mathbf{v}, \rho)$ that represents the inequality $\theta \le -(Q(\mathbf{x}_l) - l_0)(\sum_{(i,j)\in S(\mathbf{x}_l)} x_{ij} - \sum_{(i,j)\notin S(\mathbf{x}_l)} x_{ij} - |S(\mathbf{x}_l)|) - Q(\mathbf{x}_l)$. Here $S(\mathbf{x}_l) = \{(i, j) | x_{lij} = 1\}$. The constant l_0 is a lower bound of $Q(\mathbf{x})$ over $\mathbf{x} \in X$ [1]. We can set $l_0 = 0$ in our case.

To efficiently compute $Q(\mathbf{x})$, we apply the submodularity result derived in Section 4.1. More precisely, the procedure can be described as follows:

(a) Let Π be a set containing information on cancellations, and initially $\Pi = \varnothing$, $\mathbf{x}$ is a given first stage solution, $n=1$;

(b) If $n=1$, calculate its first stage realization $\hat{s}_n$ by (16), get its objective value $Q(\mathbf{x}, n)$ and cancellation decision $\mathbf{y}(n)$ by integer subproblem (24)–(28), and meanwhile store a triple $(\hat{s}_n, \hat{s}_n, \mathbf{y}(n))$ into Π, here $\hat{s}_n$ acts as both a lower bound (LB) and an upper bound (UB) of first stage realization values that lead to cancellation decision $\mathbf{y}(n)$, $n = n + 1$;

(c) If $n \le \hat{n}$, calculate first stage realization $\hat{s}_n$ by (16),
   1. if $\hat{s}_n$ falls in $[LB^\pi, UB^\pi]$ of any triple π in Π, then we directly get its optimal cancellation decision the same as $\mathbf{y}^\pi$, evaluate its objective value $Q(\mathbf{x}, n)$;
   2. otherwise, calculate its $\mathbf{y}(n)$ by integer subproblem (24)–(28) and get its objective value $Q(\mathbf{x}, n)$. If the newly calculated $\mathbf{y}(n)$ equals $\mathbf{y}^\pi$ in any triple π in Π, we update its $UB^\pi = \max\{UB^\pi, \hat{s}_n\}$, $LB^\pi = \min\{LB^\pi, \hat{s}_n\}$, otherwise add triple $(\hat{s}_n, \hat{s}_n, \mathbf{y}(n))$ to Π, let $n = n + 1$ and then go to step (c).

Our overall L-shaped algorithm follows the same general structure as the algorithm described in Angulo et al. [1] and can be found in Appendix B.1.

6. Computational results

In this section, we will apply the methods and algorithms developed in this paper to solve instances derived from Shanghai General Hospital data, and analyze how reductions of resource unavailability and exogenous cancellation can alleviate the problems caused by overcrowding. Moreover, we present comparative analysis on the ESC model which allows scheduled cancellations and the ECO model which does not. To this end, we employ the methods developed in previous sections to obtain lower and upper bounds on the performance of these models, cf. Appendix D. Before discussing the results in Section 6.3, we consider the setup of the experiments in this section.

We consider two scheduling models:

1. Exogenous Cancellations Only (ECO): Patients are scheduled a day ahead, and processed accordingly. (Exogenous cancellations still occur.)
2. Exogenous and Scheduled Cancellations (ESC): Cancellation of surgeries (in reverse order of the scheduled sequence) is allowed after the completion of the first shift (as introduced in Section 3).

The ECO model is obtained by imposing $y_{ij}(\mathbf{s}) = 0$ for all i, j in the ESC model.

Surgery time distribution: To apply the methods developed in this paper, we fit surgery time distributions to surgery data collected between October 2013 and October 2014 at Shanghai General Hospital. For practical and statistical reasons, we consider instances containing the six surgery classes with highest volumes over this period. Our tests revealed that the log-normal distribution fits the data well, as is confirmed by the QQ-plots depicted in Appendix C. The corresponding parameters are given in Table 1. Note that the flexibility of the SAA approach can easily deal with the log-normal distribution that is difficult to handle analytically.

The base case: Having estimated these surgery time distributions, we now first construct a basic problem instance, referred to as base case, and consider variations for the purpose of sensitivity analysis. For the base case, we set surgical time distributions for six patient classes based on Table 1. To account for surgery specific set-up times, we add 5 min to the surgery durations, which is close to the median reported setup time. We assume that 3 patients are available for each of the six classes, so n_p=18. On the basis of the evidence reported in Section 2, we set the probability of exogenous cancellations to 15%. Following personal communication and data analysis regarding the time between surgeries which exceeds the regular setup time, we estimate the time lost per exogenous cancellation to be 15 min. Adding 5 min of normal setup time reserved for the next patient, this gives 20 min in total to prepare the next patient in case of exogenous cancellation. Resource unavailability is also derived from Shanghai General Hospital data. We estimate the average daily resource unavailability to equal 2 h, which we divide evenly over the shifts. Specifically, we set resource unavailability for both first and second shift as i.i.d log-normal distributions with parameters $\mu = 4$ and $\sigma = 0.5$, and thus a mean of approximately 62 min. Table 2 gives the intervals on which the overtime cost function is linear, as well as the slopes for those intervals. Overtime costs are thus only incurred after regular working hours, which has a duration of 9.5 h, and additional costs for excessive overtime are occurred after 11.5 h. Lacking specific financial data, as well as data on health benefits from surgery, we normalize the reward r for each of the surgeries to equal the expected surgical duration m', where m' equals m plus the five minutes' preparation time. The penalty associated with scheduled cancellation is set to 1.05 times the reward in the base case.

Table 1
The mean (m) and standard deviation (s) (in minutes) of the log-normal distribution with parameters (μ, σ) fitted to data for various surgery classes, surgery classes are sorted in increasing order of mean.

Departments	Index	Number of observations	Log-normal parameters		Mean and std deviation	
			μ	σ	m	s
Obstetrical	1	2949	4.02	.41	60.75	25.75
Gynecology	2	5368	4.11	.88	90.14	97.62
Orthopedic	3	2236	4.70	.59	130.86	84.70
General	4	4003	4.85	.59	152.13	98.91
Thoracic	5	1303	4.98	.52	165.67	91.82
Neurosurgical	6	1234	5.06	.68	197.67	150.42

Table 2
Intervals and slopes of the overtime cost function and associated terminology.

Terminology	Regular time	Regular overtime	Excessive overtime
Interval (min)	[0,570]	(570,690]	(690, ∞)
Slope	0	1.5	2.0

6.1. Results

Section 6.1.1 investigates the performance of the developed solution methods for the base case and three variations. It also presents the comparative analysis between ESC and ECO in terms of optimal solution values. Section 6.1.2 investigates the impact of decreasing resource unavailability and reducing exogenous cancellations as means to alleviate the problems caused by overcrowding.

6.1.1. Comparative analysis

We consider four cases in order to compare the performances of the ESC and ECO policies. The three variations of the base case are obtained by varying the rewards and penalties of the surgery classes. Note that overtime costs, rewards, and cancellation costs should be understood relative to each other: the cost coefficients measure the relative importance of achieving the various conflicting objectives. The final objective is referred to as yield. Table 3 lists the variations and the base case. Remember that m' is the average surgical time including preparation time, which is set at $m' = m + 5$. For the resulting cases, we determine the yields obtained by our algorithms for ESC and ECO, as well as associated upper bounds. The results are summarized in Table 4. The table shows that, with one mild exception, our algorithm consistently finds solutions that are within 1% from the corresponding upper bound. In view of the stochasticity involved in the third stage of the three-stage recourse model, after the scheduled cancellations are decided, we consider this performance satisfactory.

Table 5 compares the solution values obtained for ESC and ECO and provides insight on the benefits of allowing scheduled cancellations. Naturally, these benefits depend on the cancellation cost. The benefit of scheduled cancellations is as much as 11.23% in the base case, and then reduces as scheduled cancellations become

Table 3
The base case and three variations for computing the reward r and cancellation penalty c from the mean m' and standard deviation s of the surgery time plus preparation time.

Case	Reward (r)	Penalty (c)
Base case	m'	1.05 m'
Case a	m'	1.2 m'
Case b	$m' + 0.5\,s$	1.05($m' + 0.5\,s$)
Case c	$m' + 0.5\,s$	1.2($m' + 0.5\,s$)

Table 4
The yield obtained by the ESC and ECO scheduling policies using the algorithms developed in this paper, as well as associated upper bounds and optimality gaps.

Policy	Statistic	Base case	Case A	Case B	Case C
	Yield	397.02 ± 0.51	376.29 ± 0.69	609.22 ± 0.69	589.21 ± 0.83
ESC	Upper bound	400.61 ± 0.65	381.15 ± 0.91	614.25 ± 2.12	594.10 ± 1.88
	Gap (%)	(0.90 ± 0.20)	(1.27 ± 0.42)	(0.82 ± 0.46)	(0.82 ± 0.45)
	Yield	356.95 ± 0.62	356.95 ± 0.62	585.47 ± 0.93	585.47 ± 0.93
ECO	Upper bound	359.86 ± 0.49	359.86 ± 0.49	585.47 ± 0.93	585.47 ± 0.93
	Gap (%)	(0.81 ± 0.31)	(0.81 ± 0.31)	(0 ± 0.32)	(0 ± 0.32)

Table 5
The improvement of ESC over ECO for each of four cases, as well as the ratio between the costs of cancelling a surgery versus the cost of performing the surgery in (excessive) overtime.

Statistic	Base case	Case A	Case B	Case C
ESC vs ECO ((ESC-ECO)/ECO, in %)	(11.23 ± 0.32)	(5.42 ± 0.37)	(4.06 ± 0.28)	(0.64 ± 0.30)
Cost ratio of cancellation vs regular overtime	1.05:1.275	1.20:1.275	1.41:1.275	1.61:1.275
Cost ratio of cancellation vs excessive overtime	1.05:1.70	1.20:1.70	1.41:1.70	1.61:1.70

penalized heavier. In comparison to a weak and simple upper bound which assumes that there is a revenue of 1 for every expected non-idle unit of regular operating room time and no cost of cancellation or overtime work, ESC closes around 50% of the gap between this bound and the solution value for ECO. A similar result holds for case B. ESC closes less than 25% of this gap for the cases A and C.

To allow the reader to appreciate the effects of increasing the costs of scheduled cancellations, we tabulate the cost ratio between performing a surgery in overtime and cancelling the surgery, as well as the ratio between performing a surgery in excessive overtime and cancelling the surgery. (Note that the cancellation decision is nontrivial even though these ratios are known: At the moment of deciding on scheduled cancellations there is considerable uncertainty regarding the starting times of second shift surgeries.) These ratios vary case by case. They also depend to a limited extent on the surgery class, but relative variation is less than 8.86% over the six surgery classes. Table 5 gives the average ratio over the six surgery classes for each case. The calculated ratios account for the probability of exogenous cancellation in the expected surgery durations. Overtime is associated with increased risks of complications and medical errors, as well as dissatisfaction among scarce staff. By consequence, hospitals may prefer scheduled cancellations and delays of the corresponding patients to the next day to performing the surgery in overtime. The *base case* assumes that it is $\frac{1.275}{1.05}$ times more desirable to cancel a surgery than to perform it in regular overtime. For the base case, the yield improvement of ESC over ECO is $(11.23 \pm 0.32)\%$, which shows that there is considerable value in allowing scheduled cancellations, even if the cancellation decision is to be taken already after completing the first shift of at most half of the scheduled patients. Cases *A* and *B* represent cases where scheduled cancellations are only $\frac{1.275}{1.20}$ and $\frac{1.275}{1.41}$ times more desirable than performing the corresponding surgeries in regular overtime, while excessive overtime is still much more undesirable relative to scheduled cancellation. In that case, the value of allowing cancellations reduces to $(5.42 \pm 0.37)\%$ and $(4.06 \pm 0.28)\%$, respectively. In Case *C*, the penalty for scheduled cancellation is so high that the recourse offers little improvement opportunity. It is therefore not surprising that the value of scheduled cancellations is very limited in case C at $(0.64 \pm 0.30)\%$.

6.1.2. *Insights*

In this section, reward and penalty cost are fixed to the base case, and we explore the impact of reducing exogenous cancellations and resource unavailability to the ESC and ECO policies as means to alleviate hospital overcrowding problems. We vary the rate of exogenous cancellations to be 0%, 5% and 15%. The latter is based on existing evidence reported in the scientific literature (cf. Section 2). The 5% appears to be a lower bound among the values reported in the scientific literature. The 0% scenario merely gives insight in the overall potential of eliminating exogenous cancellations altogether. The results are shown in Fig. 3. The figure shows that exogenous cancellation has a significant adverse impact on performance: For the ESC policy, yields increase by 5% as exogenous cancellation rate decreases from 15% to 0%. For the ECO policy, this increase is 8%. These results show that ESC can better control the adverse impact of cancellations than ECO. For all tested values of exogenous cancellation rate, the ESC policy significantly outperforms the ECO policy by about 9%, which underlines the potential value of scheduled cancellations in dealing with uncertainties, even if they will be reduced in the future.

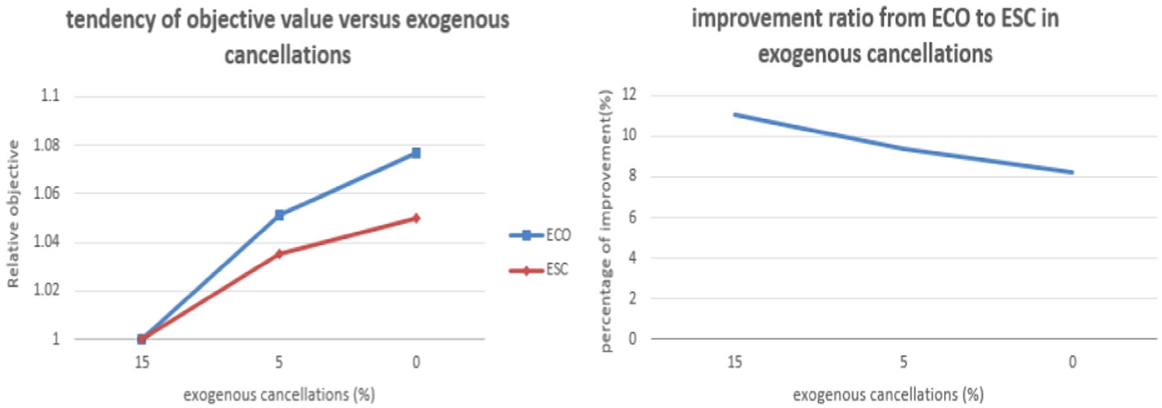

Fig. 3. The impact of the exogenous cancellations. The left figure shows the relative improvement in objective as the exogenous cancellation rate is reduced, where the improvement is measured with respect to the base case of 15% exogenous cancellation rate for both ECO and ESC. The right figure shows the relative improvement of ESC over ECO as the exogenous cancellation rate is reduced.

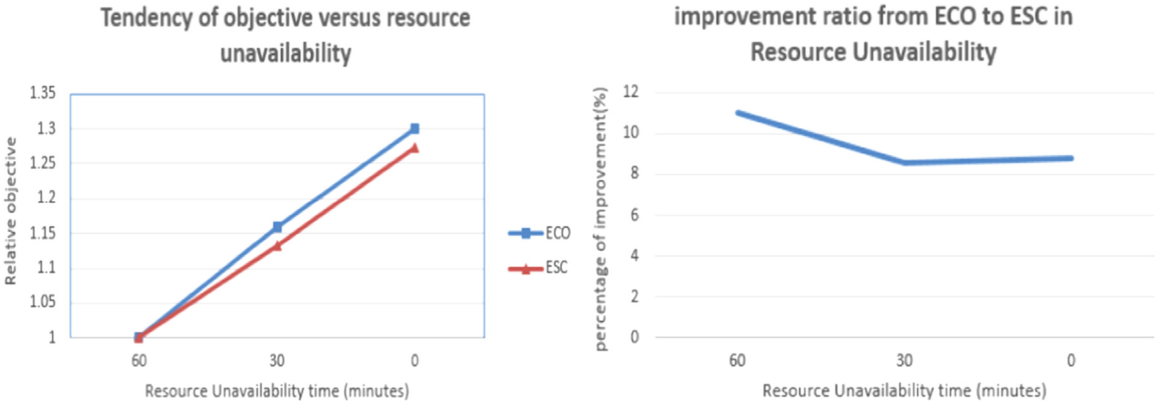

Fig. 4. The impact of the resource unavailability. The left figure shows the relative improvement in objective as the resource unavailability is reduced, where the improvement is measured with respect to the base case of resource unavailability equal 60 min for both ECO and ESC. The right figure shows the relative improvement of ESC over ECO as the resource unavailability is reduced.

Next, we investigate the sensitivity of the optimal solutions to resource unavailability. To this purpose, we vary the unavailability while keeping other parameters as in the base case. The mean resource unavailability is set to 0 (lognormal with $\mu = 0$, $\sigma = 0$), 33 (lognormal with $\mu = 3$, $\sigma = 1$) and 62 (lognormal with $\mu = 4$, $\sigma = 0.5$) minutes per shift. The latter value is derived from the provided recent data and used as benchmark. Notice that the latter value corresponds to an unavailability of slightly over two hours on a 9.5 h working day, and hence to about 21.5%. Fig. 4 shows that resource unavailability has a more than proportional adverse impact on performance: Resource unavailability of (on average) 62 min per shift reduces expected yield by around 25% for the obtained solution for ESC and even more for ECO. The ESC model significantly outperforms ECO by at least 9%, and mostly when unavailability is highest. ESC offers an increasing advantage as the unavailability increases. This is further confirmed by experiments where we compare solutions which ignore the expected resource unavailability. For the ESC model, this results in a small but highly significant decrease in solution value (of around 1%), whereas the highly significant decrease exceeds 5% for ECO.

7. Discussion and practical implications

This work considers single operating room scheduling problems as they occur in overcrowded Chinese hospitals. Overcrowding is caused by societal and economic developments which are likely to sustain for years to come. As it severely impacts access to health care, as well as the quality and safety of care when solutions are sought in working long overtime hours, adequate solution methods for these scheduling problems are urgently called for. The scheduling problems are complicated by frequent cancellations for reasons that are exogenous to operating room management, such as cancellations by patients for economic reasons, and cancellations because of recent (hospital) acquired infections. Moreover, the operating rooms suffer from human resource unavailabilities as caused by urgent demands in other departments in the overcrowded hospitals. These stochastic characteristics make the resulting scheduling problems significantly more challenging to solve than previously studied stochastic operating room scheduling problems in the scientific literature, which primarily take stochastic surgery times into account.

Our study analyzes the impact of the exogenous cancellations and resource unavailabilities on the optimal schedules, so as to understand if and how reducing the exogenous cancellations and resource unavailabilities can assist hospitals to cope with the sustained excess demand. To this purpose, we developed solution methods for the presented operating room scheduling problems. Moreover, we analyzed the known practice of scheduled cancellations, which from a modelling perspective defines a second stage recourse moment in the stochastic scheduling problem.

The resulting problem forms a three-stage scheduling problem with recourse, as the realizations of the exogenous cancellations, unavailability and surgery durations for a second shift of patients only become known after the second stage decisions on scheduled cancellations have been made. We solve the three-stage recourse problem using sample average approximation methods and corresponding optimization techniques. Because of the stochasticity involved in the third stage however, the lower and upper bounds available are slightly weaker than it is often the case in two stage problems, and computation times can become larger. To remedy these computational problems, we derive several structural properties on the optimal schedule and scheduled cancellations, which allow us to speed up the optimization. Thus the proposed sample approximation approach which relies on the L-shaped method and optimality cuts forms a nontrivial innovation in stochastic scheduling itself. The developed solution methods deliver solutions which are mostly within 1% of optimal, thus allowing comparative analysis and sensitivity analysis of the various scheduling models by considering their solutions.

In many current practices, operating room schedules are composed without explicit consideration of the stochastic processes involved (yet only considering mean surgery times), or even without evaluation of the schedule at all. Our research firstly shows that the stochasticity of human resource unavailability, exogenous cancellations and procedure times can be simultaneously included in a scheduling model, for which good quality solutions balancing overtime costs with high workloads can be found. Our results show that taking the stochasticity into account yields substantially and significantly better operating room schedules. The improvements obtained for solutions with scheduled cancellations of up to 11% are much above the upper bounds on the solutions without scheduled cancellations, thus ensuring that the optimality gaps do not invalidate the conclusions.

Implementing the approach may take prolonged effort because substantial data collection is needed. But our results indicate that significant and substantial improvements are already attainable by (a) taking unavailability and no-show explicitly into account when constructing the initial schedules, and (b) systematic use of (early) scheduled cancellations. Likely benefits are better control of operating costs, increased staff satisfaction, and improvement of patient safety and satisfaction.

With these solution methods at hand, we have further analyzed exogenous cancellation and resource unavailability. As the latter may be in the order of 20% of regular opening hours, it is clear that improving unavailability holds great potential to alleviate the problems caused by overcrowding. Our results reveal that – while scheduled cancellations can limit the negative impact of resource unavailability – the overall impact on the solution values is more than proportional to the unavailability and exceeds 25% in the presented instances. A practical implication is therefore that hospitals and patients can greatly benefit from better management and control of the operational deployment of (human) resources to reduce their unavailability.

Although evidence indicates that exogenous cancellation may apply to as much as 15% of scheduled surgeries, it poses fewer difficulties for operating room scheduling and utilization than resource availability. This holds particularly true for the ESC model as its optimal solution value does not improve beyond 5%, even when exogenous cancellation is reduced by the full 15%. The impact for the model without scheduled cancellations is larger, confirming the potential of scheduled cancellations. From a practical operating room management perspective, these results imply that reduction of exogenous cancellations is worth considering after implementation of scheduled cancellations and reducing human resource unavailability. Especially so as the causes of scheduled cancellations are beyond the control of operating room management. As exogenous cancellations often follow from financial barriers and worsening of health status, reducing exogenous cancellations remains of urgency and importance.

While our analysis relies on data from a single hospital, Shanghai General Hospital, we believe that the model, solution methods, and analyses are likely to have relevance for the many other level 3 large city hospitals in China, which are presently overcrowded and face further demand increases. Similar problems occur in other developing countries as well. Our research presents first theoretical advancements on the resulting operating room scheduling problems as well as practical improvement suggestions. At the same time, it is clear that it has limitations and poses new research questions. For example, models which set the recourse moment at a fixed moment in time, or divide the shifts based on minutes workload rather than numbers of patients are worthy of further study. Moreover, one may consider the problem of determining the optimal moment in time, workload minutes, or relative patient number after which to end the first shift. We therefore hope that our research motivates other researchers to advance the work on the presently under-researched urgent operations management problems occurring in the operating rooms of China and other – mostly developing – countries, serving the far majority of the global population.

Appendix A. Proofs of lemmas and theorems

Lemma 1. *The second stage decision problems* $Q(\mathbf{x}, \mathbf{s})$ *and* $\tilde{Q}(\mathbf{x}, \mathbf{s})$ *are equivalent for any* $\mathbf{x} \in X$ *and realization of* $\mathbf{s}$.

Proof. Let $y_{ij}(\mathbf{s})$ and $k_{\mathbf{s}}$ are, respectively, the optimal solution to $Q(\mathbf{x}, \mathbf{s})$ and $\tilde{Q}(\mathbf{x}, \mathbf{s})$. Then $y_{ij}(\mathbf{s}) \in Y(\mathbf{x})$ and $k_{\mathbf{s}} \in Z$.

- Let k satisfy $\sum_{i=1}^{n_p}\sum_{j=1}^{k} x_{ij} = k_{\mathbf{s}}$, next we equivalently transform $k_{\mathbf{s}}$ into a solution $y_{ij}^{k}(\mathbf{s})$:

$$y_{ij}^{k}(\mathbf{s}) = \begin{cases} 0, & \forall\ i = 1, \ldots, n_p,\ \ j = 1, \ldots, k, \\ x_{ij}, & \text{otherwise} \end{cases} \tag{A.1}$$

  Obviously, $y_{ij}^{k}(\mathbf{s})$ is a feasible solution to $Q(\mathbf{x}, \mathbf{s})$, and $\tilde{Q}(\mathbf{x}, \mathbf{s}) \ge Q(\mathbf{x}, \mathbf{s})$.

- Let

$$k_{\mathbf{s}}^{y} = \sum_{i=1}^{n_p}\sum_{j=1}^{\kappa}(x_{ij} - y_{ij}(\mathbf{s})) \tag{A.2}$$

  then $k_{\mathbf{s}}^{y} \in Z$ and $\tilde{Q}(\mathbf{x}, \mathbf{s}) \le Q(\mathbf{x}, \mathbf{s})$.

Summarizing above, we can conclude that $Q(\mathbf{x}, \mathbf{s})$ and $\tilde{Q}(\mathbf{x}, \mathbf{s})$ are equivalent for any $\mathbf{x} \in X$ and realization of $\mathbf{s}$.□

Proposition 1. *Let* $\mathbf{x} \in X$ *be given, and condition on* $\hat{s}$, *then* $g(\hat{s}, k_{\hat{s}})$ *is a supermodular function.*

Proof. To prove that $g(\hat{s}, k_{\hat{s}})$ is supermodular in $(\hat{s}, k_{\hat{s}})$, we should prove that $\forall\ \hat{s}_1 \ge \hat{s}_2,\ k_{\hat{s}_1} \le k_{\hat{s}_2}$,

$$g(\hat{s}_1, k_{\hat{s}_2}) + g(\hat{s}_2, k_{\hat{s}_1}) \ge g(\hat{s}_1, k_{\hat{s}_1}) + g(\hat{s}_2, k_{\hat{s}_2}) \tag{A.3}$$

Expanding their expressions and merge similar items, inequality (A.3) is equivalent to the following:

$$E_{\xi}D\left(\sum_{i=1}^{n_p}\sum_{j=1}^{k_{\hat{s}_2}} x_{ij}\xi_i + \hat{s}_1 + \eta_2\right) + E_{\xi}D\left(\sum_{i=1}^{n_p}\sum_{j=1}^{k_{\hat{s}_1}} x_{ij}\xi_i + \hat{s}_2 + \eta_2\right)$$
$$\ge E_{\xi}D\left(\sum_{i=1}^{n_p}\sum_{j=1}^{k_{\hat{s}_1}} x_{ij}\xi_i + \hat{s}_1 + \eta_2\right) + E_{\xi}D\left(\sum_{i=1}^{n_p}\sum_{j=1}^{k_{\hat{s}_2}} x_{ij}\xi_i + \hat{s}_2 + \eta_2\right) \tag{A.4}$$

The convexity of function $D(x)$ in x justifies inequality (A.4) and the proof is done.□

Proposition 2. *Let* $\mathbf{x} \in X$ *be given, and consider two realizations of the total time for the first shift:* $\hat{s}_1$ *and* $\hat{s}_2$ *with* $\hat{s}_1 \le \hat{s}_2$, *then* $k_{\hat{s}_1}^{*} \ge k_{\hat{s}_2}^{*}$.

Proof. As Proposition 1 showed, $g(\hat{s}, k_{\hat{s}})$ is supermodular in vector $(\hat{s}, k_{\hat{s}})$, by introducing $t = -\hat{s}$, we can get submodular function $g(t, k_{\hat{s}})$, and applying the property of submodular function [43], we can get that $k_{\hat{s}}^{*}$ increases in t, i.e., $k_{\hat{s}}^{*}$ decreases in $\hat{s}$.□

Lemma 2. *Let* $\mathbf{x} \in \bar{X}$ *be given and* $\hat{s}$ *be defined by (17), then* $F(\hat{s}, \mathbf{y})$ *is convex in* $\mathbf{y} \in \bar{Y}(\mathbf{x})$.

Proof. We will prove that $\forall\ \mathbf{y}_1 = (y_{ij}^{1})_{n_p\times\kappa},\ \mathbf{y}_2 = (y_{ij}^{2})_{n_p\times\kappa}$ and $\lambda \ge 0$,

$$F(\hat{s}, \lambda\mathbf{y}_1 + (1-\lambda)\mathbf{y}_2) \le \lambda F(\hat{s}, \mathbf{y}_1) + (1-\lambda)F(\hat{s}, \mathbf{y}_2)$$

remark that $\sum_{i=1}^{n_p}\sum_{j=1}^{\kappa} I_0(\xi_i)r_i(x_{ij} - y_{ij})$ is linear in $\mathbf{y}$ and make no difference in the convexity, the above inequality holds if

$$E_{\xi}D\left[\sum_{i=1}^{n_p}\sum_{j=1}^{\kappa}\xi_i[\lambda(x_{ij} - y_{ij}^{1}) + (1-\lambda)(x_{ij} - y_{ij}^{2})] + \hat{s} + \eta_2\right]$$
$$\le \lambda E_{\xi}D\left[\sum_{i=1}^{n_p}\sum_{j=1}^{\kappa}\xi_i(x_{ij} - y_{ij}^{1}) + \hat{s} + \eta_2\right]$$
$$+ (1-\lambda)E_{\xi}D\left[\sum_{i=1}^{n_p}\sum_{j=1}^{\kappa}\xi_i(x_{ij} - y_{ij}^{2}) + \hat{s} + \eta_2\right]$$

Since $D(x)$ is convex in x, the second inequality holds for any $\lambda \in [0, 1]$ and $\mathbf{y}_1, \mathbf{y}_2$, and the proposition is true.□

Proposition 3. *Let* $\mathbf{x} \in X$ *be given and* $\hat{s}$ *be defined by (17), then* $f(\hat{s}) = \min_{\mathbf{y}\in\bar{Y}} F(\hat{s}, \mathbf{y})$ *and* $f(\hat{s})$ *is convex in* $\hat{s}$. *Besides,* $R(\mathbf{s})$ *defined in* $\sum_{i=1}^{n_p} I_0(s_i)r_i x_{i0}$ *is also convex in* $\mathbf{s}$.

Proof. We can easily get $f(\hat{s}) = \min_{\mathbf{y}\in\bar{Y}} F(\hat{s}, \mathbf{y})$ by Lemma 1. Next we will prove that $\forall\ \hat{s}_1, \hat{s}_2,\ \hat{s}_1 \ge \hat{s}_2$ and $\lambda \ge 0$,

$$f(\lambda\hat{s}_1 + (1-\lambda)\hat{s}_2) \le \lambda f(\hat{s}_1) + (1-\lambda)f(\hat{s}_2) \tag{A.5}$$

Let $\mathbf{y}_1 \in \bar{Y}$ and $\mathbf{y}_2 \in \bar{Y}$ be, respectively, the optimal solution to $f(\hat{s}_1)$

and $f(\hat{s}_2)$, then $\lambda \mathbf{y}_1 + (1 - \lambda)\mathbf{y}_2$ is a feasible solution to $f(\lambda\hat{s}_1 + (1 - \lambda)\hat{s}_2)$ and

$$f(\lambda\hat{s}_1 + (1 - \lambda)\hat{s}_2) \le F(\lambda\hat{s}_1 + (1 - \lambda)\hat{s}_2, \lambda\mathbf{y}_1 + (1 - \lambda)\mathbf{y}_2)$$

What's more,

$$\begin{aligned}
&\lambda f(\hat{s}_1) + (1 - \lambda)f(\hat{s}_2) - F(\lambda\hat{s}_1 + (1 - \lambda)\hat{s}_2, \lambda\mathbf{y}_1 + (1-\lambda)\mathbf{y}_2)\\
&= \lambda E_{\xi}[D\left(\sum_{i=1}^{n_p}\sum_{j=1}^{\kappa} \xi_i(x_{ij} - y_{ij}^1) + \hat{s}_1 + \eta_2\right) + \sum_{i=1}^{n_p}\sum_{j=1}^{\kappa} I_0(\xi_i)r_i(x_{ij} - y_{ij}^1)]\\
&\quad + (1 - \lambda)E_{\xi}[D\left(\sum_{i=1}^{n_p}\sum_{j=1}^{\kappa} \xi_i(x_{ij} - y_{ij}^2) + \hat{s}_2 + \eta_2\right)\\
&\quad + \sum_{i=1}^{n_p}\sum_{j=1}^{\kappa} I_0(\xi_i)r_i(x_{ij} - y_{ij}^2)]\\
&\quad - E_{\xi}[D\left(\sum_{i=1}^{n_p}\sum_{j=1}^{\kappa} \xi_i(x_{ij} - \lambda y_{ij}^1 - (1 - \lambda)y_{ij}^2\right) + \lambda\hat{s}_1 + (1 - \lambda)\hat{s}_2 + \eta_2)\\
&\quad + \sum_{i=1}^{n_p}\sum_{j=1}^{\kappa} I_0(\xi_i)r_i(x_{ij} - \lambda y_{ij}^1 - (1 - \lambda)y_{ij}^2)]\\
&= \lambda E_{\xi}D\left[\sum_{i=1}^{n_p}\sum_{j=1}^{\kappa} \xi_i(x_{ij} - y_{ij}^1) + \hat{s}_1 + \eta_2\right]\\
&\quad + (1 - \lambda)E_{\xi}D\left[\sum_{i=1}^{n_p}\sum_{j=1}^{\kappa} \xi_i(x_{ij} - y_{ij}^2) + \hat{s}_2 + \eta_2\right]\\
&\quad - E_{\xi}D\left[\lambda\left(\sum_{i=1}^{n_p}\sum_{j=1}^{\kappa} \xi_i(x_{ij} - y_{ij}^1) + \hat{s}_1 + \eta_2\right)\right.\\
&\quad \left. + (1 - \lambda)\left(\sum_{i=1}^{n_p}\sum_{j=1}^{\kappa} \xi_i(x_{ij} - y_{ij}^2) + \hat{s}_2 + \eta_2\right)\right] \ge 0
\end{aligned}$$

Since $D(x)$ is convex in x, the last inequality holds and inequality (A.5) is true. Moreover, $I_0(s_i)$ is convex in $\mathbf{s}$, which can directly derive that $R(\mathbf{s})$ is also convex in $\mathbf{s}$.□

Appendix B. Application of L-shaped method

B.1. Algorithm of L-shaped method

Our overall L-shaped algorithm follows the same general structure as the algorithm described in Angulo et al. [1]. Based on the above analysis, it can be described as follows:

Algorithm 1.

Step 0 Initiate $\Theta = \varnothing$. Throughout, $\boldsymbol{\Theta}$ will be used for the master problem.

Step 1 Optimize the integral master problem to obtain an optimal solution $\mathbf{x} \in X$ and corresponding objective value z and θ. If $\frac{\theta - (-Q_{LP}(\mathbf{x}))}{(-Q_{LP}(\mathbf{x}))} > \epsilon$, add the corresponding optimality cut $(\mathbf{v}, \rho)$ to $\boldsymbol{\Theta}$ and go to Step 1, otherwise go to Step 2.

Step 2 If $\frac{\theta - (-Q(\mathbf{x}))}{(-Q(\mathbf{x}))} > \epsilon$, add the corresponding integer optimality cut to $\boldsymbol{\Theta}$, and go to Step 1. Otherwise, if $\frac{\theta - (-Q(\mathbf{x}))}{(-Q(\mathbf{x})}) \le \epsilon$, terminate, designating $\mathbf{x}$ as the ϵ-optimal solution.

B.2. Upper bound by Jensen's inequality

The L-shaped method from the previous section can be enhanced by adding Jensen's inequality as a constraint to both the integral and relaxed master problem. By the results obtained in Section 4.2, the second stage costs can be bounded from below if all first shift surgeries take on their expected value (for the SAA approach, this translated to replacing the expected value by the sample mean). To explicitly give the constraints, let $\bar{s}_i = \frac{\sum_{n=1}^{\hat{n}}(s_{in})}{\hat{n}}$, and $\bar{\eta}_1 = \frac{\sum_{n=1}^{\hat{n}}\eta_{1n}}{\hat{n}}$, and let $\bar{\mathbf{y}} = \{\bar{y}_{ij}|i \in I, j \in \{1, \dots, \kappa\}\}$ denote the second stage decisions if all first-stage random variables take on their expected value, in which case $I_0(\bar{s}_i) = 0$. Then,

$$\theta \le -\sum_{i=1}^{n_p}\sum_{j=1}^{\kappa} c_i\bar{y}_{ij} - \frac{1}{\hat{m}}\sum_{m=1}^{\hat{m}}\left[\sum_{v=0}^{q}\tau_v\bar{\phi}_v(m) + \sum_{i=1}^{n_p}\sum_{j=1}^{\kappa} I_0(\xi_{im})r_i(x_{ij} - \bar{y}_{ij})\right] \tag{B.1}$$

$$\sum_{v=0}^{q}\bar{\phi}_v(m) = \sum_{i=1}^{n_p}\sum_{j=1}^{\kappa}\xi_{im}(x_{ij} - \bar{y}_{ij}) + \eta_{2\,m} + \sum_{i=1}^{n_p}\bar{s}_i x_{i0} + \bar{\eta}_1 \tag{B.2}$$

$$\bar{\mathbf{y}} \in \bar{Y}(\mathbf{x}) \tag{B.3}$$

$$\bar{\varphi}_v(m) \in [0, l_v], \quad m \in \{1, \dots, \hat{m}\}, \quad v \in \{0, \dots, q\} \tag{B.4}$$

These constraints extend the results in Batun et al. [3] for our problem.

Appendix C. QQ-plots of surgery time distribution

The log-normal distribution fits the surgery time data quite well, as is confirmed by the QQ-plots depicted in Fig. C1.

Appendix D. Obtaining performance bounds

The general method for obtaining upper and lower bound estimates from the SAA of two-stage stochastic programs has been discussed in Mak et al. [30] and Kleywegt et al. [21]. Let us recall, however, that the ESC scheduling problem is a three-stage SP. Upper and lower bounds are therefore obtained from the SAA (23)–(28) as follows. (Recall that we are maximizing.) The SAA objective averages all combinations of $\hat{n}$ first shift samples and $\hat{m}$ second shift samples, which equals a total of $\hat{n} \times \hat{m}$ combinations. Because the problem is three-stage, it requires relatively many samples to sufficiently accurately represent the randomness. In our numerical experiments we use $\hat{n} = \hat{m} = 500$, for a total of 250,000 combinations. The target accuracy ϵ of Algorithm 1 is set at 0.5% when running time is shorter than 24 h, and is increased to 2% when this running time bound is exceeded. An upper bound estimate is obtained by averaging the upper bound on the optimal objective value for 10 collections of $\hat{n} \times \hat{m}$ samples. (Thus, the our upper bound becomes weaker as ϵ increases.)

To obtain a lower bound estimate, we select a solution $\mathbf{x}' \in X$ that optimizes the SAA for a 500×500 sample. We then fix the schedule to this $\mathbf{x}'$, and solve the SAA for a single first shift sample ($\hat{n} = 1$), while setting $\hat{m} = 2000$. This yields a single appropriate cancellation decision for that first shift realization. The outcome for the first shift sample with that cancellation decision is evaluated using a new, independent set of 2000 second shift realizations. This yields an unbiased lower bound estimate. A reliable lower bound estimate with associated standard deviation is obtained by averaging the result of this procedure for 2000 replications, i.e., (1) generate new first shift and second shift

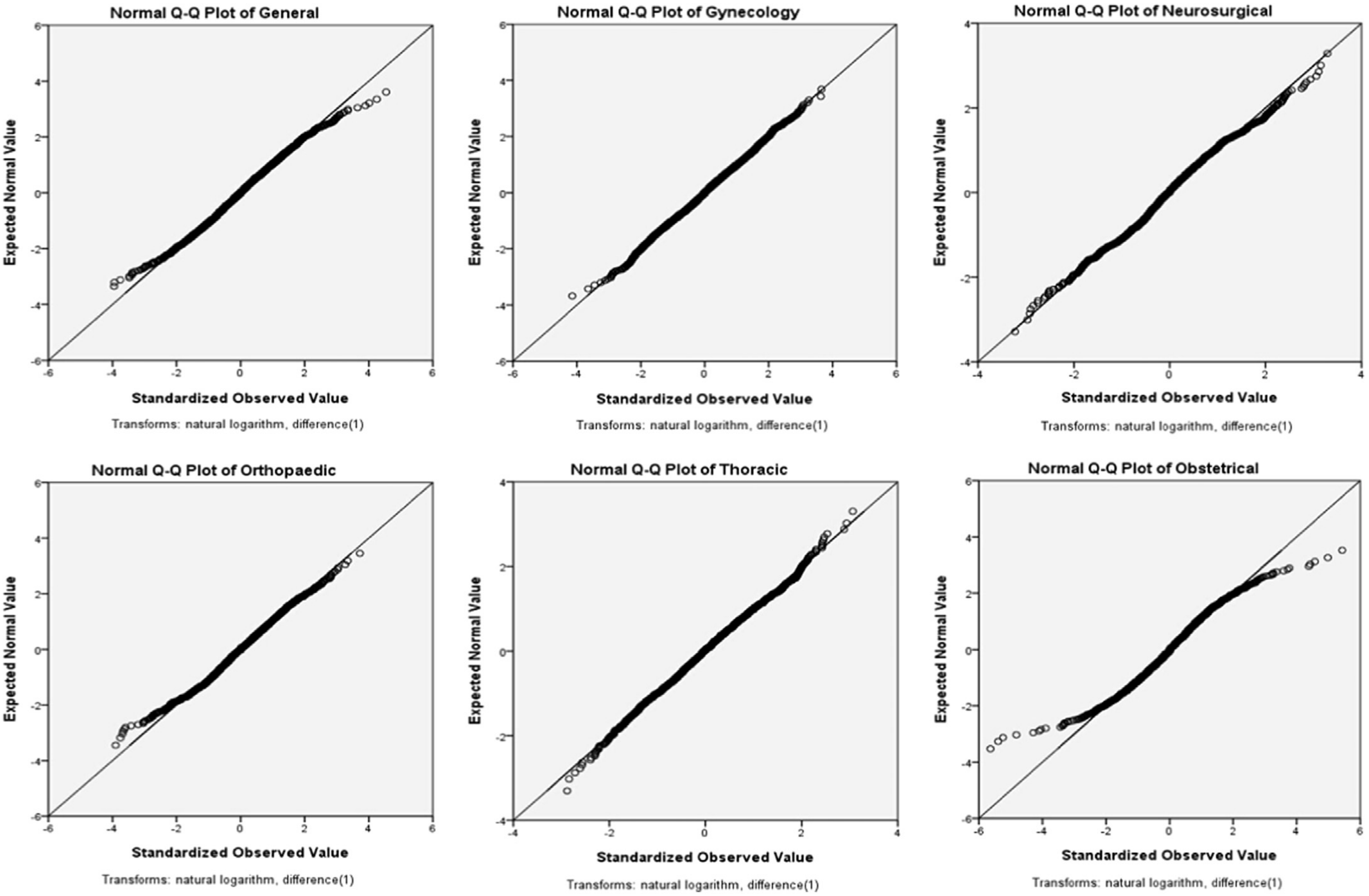

Fig. C1. QQ plots of the data versus the fitted lognormal distribution, for various surgery classes.

samples, (2) determine an appropriate cancellation decision, and (3) evaluate the outcome of the first shift sample and cancellation decision with a new second shift sample.

References

[1] Angulo G, Ahmed S, Dey SS. Improving the integer L-shaped method, Unpublished results; 2014.

[2] Barber LS, Yao L. Health insurance systems in China: a briefing note. World health report 2010 background paper 37. World Health Organization; 2010. URL ⟨http://www.who.int/healthsystems/topics/fiancing/healthreport/whr_background/en⟩.

[3] Batun S, Denton BT, Huschka TR, Schaefer AJ. Operating room pooling and parallel surgery processing under uncertainty. INFORMS J Comput 2011;23(2):220–37.

[4] Berg BP, Denton BT, Erdogan SA, Rohleder T, Huschka T. Optimal booking and scheduling in outpatient procedure centers. Comput Oper Res 2014;50:24–37.

[5] Birge JR. Decomposition and partitioning methods for multistage stochastic linear programs. Oper Res 1985;33(5):989–1007.

[6] Birge JR, Louveaux F. Introduction to stochastic programming. Berlin/Heidelberg: Springer Science & Business Media; 2011.

[7] Birge JR, Rosa CH. Parallel decomposition of large-scale stochastic nonlinear programs. Ann Oper Res 1996;64(1):39–65.

[8] Carøe CC, Tind J. L-shaped decomposition of two-stage stochastic programs with integer recourse. Math Program 1998;83(1–3):451–64.

[9] Chalya P, Gilyoma J, Mabula J, Simbila S, Ngayomela I, Chandika A, Mahalu W. Incidence, causes and pattern of cancellation of elective surgical operations in a University Teaching Hospital in the Lake Zone, Tanzania. Afr Health Sci 2011;11(3).

[10] Chiu C, Lee A, Chui P. Cancellation of elective operations on the day of intended surgery in a Hong Kong hospital: point prevalence and reasons. Hong Kong Med J 2012;18(1):5–10.

[11] Daily Briefing. Same-day surgery cancellations cost hospitals millions. URL ⟨www.advisory.com/daily-briefing/2012/05/09/same-day-surgery-cancellations-cost-hospitals-millions⟩; 2012.

[12] Denton B, Viapiano J, Vogl A. Optimization of surgery sequencing and scheduling decisions under uncertainty. Health Care Manag Sci 2007;10(1):13–24.

[13] Feldman J, Liu N, Topaloglu H, Ziya S. Appointment scheduling under patient preference and no-show behavior. Oper Res 2014;62(4):794–811.

[14] Fernando BS, Cannon PS, Mohan M. Cancellation of surgical day cases in an ophthalmic centre. Acta Ophthalmol 2009;87(3):357–8.

[15] Gade D, Küçükyavuz S, Sen S. Decomposition algorithms with parametric Gomory cuts for two-stage stochastic integer programs. Math Program 2014;144(1–2):39–64.

[16] Hussain AM, Khan FA. Anaesthetic reasons for cancellation of elective surgical inpatients on the day of surgery in a teaching hospital. J Pak Med Assoc 2005;55(9):374.

[17] Jensen JLWV. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. Acta Math 1906;30(1):175–93.

[18] Jiang D, Zhu S, Zhang X, Zhou X, Li L. Cause for selective operation cancellation in paediatrics and suggestions. Today Nurse 2011;9 [in Chinese]. URL ⟨http://d.g.wanfangdata.com.cn/Periodical_ddhs201109013.aspx⟩.

[19] Jie X, Chang H, Zhong A, Liu X, Cai Y. Causes of canceling selective operation and the management countermeasures. J Nurs Adm 2012;5:P366–7 [in Chinese]. URL ⟨http://d.g.wanfangdata.com.cn/Periodical_hlglzz201205026.aspx⟩.

[20] Karaesmen I, Van Ryzin G. Overbooking with substitutable inventory classes. Oper Res 2004;52(1):83–104.

[21] Kleywegt AJ, Shapiro A, Homem-de Mello T. The sample average approximation method for stochastic discrete optimization. SIAM J Optim 2002;12(2):479–502.

[22] Kolawole I, Bolaji B. Reasons for cancellation of elective surgery in Ilorin. Niger J Surg Res 2002;4(1):28–33.

[23] Kumar R, Gandhi R. Reasons for cancellation of operation on the day of intended surgery in a multidisciplinary 500 bedded hospital. J Anaesthesiol Clin Pharmacol 2012;28(1):66.

[24] LaGanga LR, Lawrence SR. Clinic overbooking to improve patient access and increase provider productivity. Decis Sci 2007;38(2):251–76.

[25] Lai K-K, Ng W-L. A stochastic approach to hotel revenue optimization. Comput Oper Res 2005;32(5):1059–72.

[26] Laporte G, Louveaux FV. The integer L-shaped method for stochastic integer programs with complete recourse. Oper Res Lett 1993;13(3):133–42.

[27] Li C, Jiang Z, An M, Luo D. Investigation and analysis of 560 cases of elective operation cancellation. Chin Gen Nurs 2011:1672–888 [in Chinese]. URL ⟨http://d.g.wanfangdata.com.cn/Periodical_jths201126059.aspx⟩.

[28] Liang L, Langenbrunner JC. The long march to universal coverage: lessons from

China. Worldbank report. World Bank. URL ⟨http://documents.worldbank.org/curated/en/2013/01/17207313/long-march-universal-coverage-lessons-china⟩; 2013.

[29] Long Q, Xu L, Bekedam H, Tang S. Changes in health expenditures in China in 2000s: has the health system reform improved affordability. Int J Equity Health 2013;12(1):40.

[30] Mak W-K, Morton DP, Wood RK. Monte Carlo bounding techniques for determining solution quality in stochastic programs. Oper Res Lett 1999;24(1):47–56.

[31] Mancilla C, Storer RH. Stochastic sequencing and scheduling of an operating room [Theses and dissertations]. Lehigh University, Department of Industrial and Systems Engineering; 2009.

[32] May JH, Spangler WE, Strum DP, Vargas LG. The surgical scheduling problem: current research and future opportunities. Prod Oper Manag 2011;20(3):392–405.

[33] Meara JG, Leather AJM, Hagander L, Blake CA, Nivaldo A, Emmanuel AA, et al. Global surgery 2030: evidence and solutions for achieving health, welfare, and economic development. The Lancet 2015;386 (9993) 569-624, http://dx.doi.org/10.1016/S0140-6736(15)60160-X.

[34] Meng Q, Xu L, Zhang Y, Qian J, Cai M, Xin Y, Gao J, Xu K, Boerma JT, Barber SL. Trends in access to health services and financial protection in China between 2003 and 2011: a cross-sectional study. Lancet 2012;379(9818):805–14.

[35] National Bureau of Statistics of China. China statistical yearbook. URL ⟨http://www.stats.gov.cn/tjsj/ndsj/2011/indexeh.htm⟩; 2011.

[36] National Bureau of Statistics of China. China statistical yearbook. URL ⟨http://www.stats.gov.cn/tjsj/ndsj/2013/indexeh.htm⟩; 2013.

[37] Perroca MG, Jericó MdC, Facundin SD. Surgery cancelling at a teaching hospital: implications for cost management. Rev Lat-Am De Enferm 2007;15(5):1018–1024.

[38] Shi H, Li Y. Inter-organizational service delivery in Chinese hospital industry: a social exchange perspective. Can Soc Sci 2014;10:63–71.

[39] Stepaniak P, van der Velden R, van de Klundert J, Wagelmans A. Human and artificial scheduling system for operating rooms. In: Handbook of healthcare system scheduling. Springer; Berlin/Heidelberg, 2012. p. 155–75.

[40] Subramanian J, Stidham Jr. S, Lautenbacher CJ. Airline yield management with overbooking, cancellations, and no-shows. Transp Sci 1999;33(2):147–67.

[41] Sussmuth-Syckerhoff, Wang. China's health care reforms. Mckinsey report, McKinsey. URL ⟨http://www.mckinsey.com/search.aspx?q=China%27s+health+care+reforms⟩; June 2012.

[42] Talluri KT, Van Ryzin GJ. The theory and practice of revenue management, vol. 68. Berlin/Heidelberg: Springer Science & Business Media; 2006.

[43] Topkis DM. Supermodularity and complementarity.New york: Princeton University Press; 1998.

[44] Van Slyke RM, Wets R. L-shaped linear programs with applications to optimal control and stochastic programming. SIAM J Appl Math 1969;17(4):638–63.

[45] World Bank. World Bank national accounts data. URL ⟨http://data.worldbank.org/indicator/NY.GDP.MKTP.CD⟩; 2015.

[46] Xiang M, Liu Y, Zhong J, Fan Y, Ni W. Causes for cancellation of inpatients elective operations. Hosp Adm J Chin PLA 2014;8:729–32 [in Chinese]. URL ⟨http://ss.zhizhen.com/detail_38502727e7500f261a104495da8e262a923ae0230a50f6621921b0a3ea255101fc1cf1fbb4666ae6b9afde42086c32a54027a5fd22c20becd210c55d06559d120b6c96ad4888324d18c6eceb52c3ed30⟩.

[47] Xiao G, Van Jaarsveld W, Dong M, Van de Klundert J. The adaptive operating room schedule with committing: an application of the knapsack problem. Unpublished results; 2015.

[48] Xu B, Fang C, Du L. Cause for selective operation cancellation. Anhui Med Pharm J 2009 [in Chinese]. URL ⟨http://d.g.wanfangdata.com.cn/Periodical_ahyy200901022.aspx⟩.

[49] Yoon S, Lee S, Lee H, Lim H, Yoon S, Chang S. The effect of increasing operating room capacity on day-of-surgery cancellation. Anaesth Intensive Care 2009;37(2):261–6.

[50] Zhang Y, Dai Y, Ma H, Shui Z. Analysis of the cause for day-surgery cancellation and its countermeasures. West China Med J 2014 [in Chinese]. URL ⟨http://www.cnki.net/KCMS/detail/detail.aspx?dbname=cjfd2014&filename=hxyx201410053&dbcode=CJFQ&urlid=&yx=&v=MDU1NTZyV00xRnJDVVJMK2ZZT1pvRnlubFVyL0lEUlh5VnJHNEg5WE5yNDlBWjRSOGVYMUx1eFlTN0RoMVQzcVQ=⟩.