IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING

Optimal Appointment Rule Design in an Outpatient Department

Jie Song, *Senior Member, IEEE*, Yaqing Bai, and Jianpei Wen

***Abstract*—Hospitals use appointment systems to manage patient access. Appointment rule, which consists of the length of the booking window, block capacity, and block service time, is critical to achieve efficiency and timely access to healthcare delivery. In this paper, we use a renewal process model to evaluate interday appointment planning and design improved appointment rules for hospitals, especially for those with limited or insufficient resources. We present an associated embedded Markov chain to derive the steady-state distribution. To balance the waiting time and probability of healthcare access, we propose three performance measures, namely, slot utilization, appointment success rate, and patient waiting time, for our evaluation. We then conduct a numerical study to examine the impact of each appointment rule parameter. Qualitative results show that extending the booking window does not significantly reduce system congestion, and a narrowed appointment block is a suitable design for highly in-demand doctors. We use our model and method to design an optimal appointment rule for an actual hospital in Beijing, China. The improved appointment rule is practical and useful for the decision-making of hospital managers.**

***Note to Practitioners*—An appointment system with a limited booking window length is a practical solution to reduce waiting time. However, potential patients will lose the opportunity to obtain access to healthcare systems, especially those with a high arrival rate. To balance the tradeoff between shortening the waiting time and increasing healthcare access probability, we attempt to evaluate and design an improved appointment rule that includes booking window length, block capacity, and block service time. Sensitivity analysis shows that extending the booking window does not significantly reduce system congestion, and a narrowed appointment block is a suitable design for highly in-demand doctors. On the basis of our model, we design an optimal appointment rule for an actual Chinese hospital. Results show that improvement can be significant (more than 60%, for example) depending on the parameters.**

***Index Terms*—Appointment system, congestion, embedded Markov chain, interday scheduling, patient cancellation.**

Manuscript received September 3, 2017; revised January 4, 2018; accepted January 11, 2018. This paper was recommended for publication by Associate Editor G. Q. Huang and Editor J. Li upon evaluation of the reviewers’ comments. This work was supported by the National Science Foundation of China under Grant 71671005, Grant 71731006, and Grant 71301003. *(Corresponding author: Jie Song.)*

J. Song and J. Wen are with the Department of Industrial Engineering and Management, Peking University, Beijing 100871, China (e-mail: songjie@coe.pku.edu.cn; wenjianpei@pku.edu.cn).

Y. Bai is with the Advertising Business Department, Amazon ( CN), Beijing 100025, China (e-mail: yaqinbai@amazon.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASE.2018.2794335

I. INTRODUCTION

**D**

IFFICULTY in obtaining a timely appointment is a problem in healthcare systems worldwide. Merritt Hawkins[[1]](#footnote-1)reported that patients wait 24 days on the average in 15 of the largest cities in the U.S. for a specialists appointment, and the delay for specialty care appointment, such as with a dermatologist in Boston, even takes as long as 52.4 days. This long waiting time, on the one hand, causes patients with potentially urgent problems to miss the opportunity to receive timely treatment. On the other hand, this long waiting time may lead to patient cancellation and even no-show [1] , which introduce additional uncertainty and can cause severe inefficiencies.

To shorten the waiting time of patients, several providers attempt to limit the length of the booking window (a period that can be used to make an appointment). For example, Peking University Third Hospital ( PUTH), 2 which is one of the best general hospitals in China and handles more than 4 million outpatient patients annually, has a booking window of only seven days. A short booking window can shorten a patients *indirect waiting time*, which is the difference between the time that a patient obtains an appointment and the time of being seen by a doctor. However, potential patients will lose the opportunity to obtain access to healthcare systems, especially in places where healthcare resources are limited, if all appointment slots on a particular day are fully booked. This scenario leads to a long *potential waiting time*, which is the period starting from the time that a patient subjectively wishes to make an appointment until the time that he/she obtains it successfully.

Designing an effective appointment system involves a tradeoff between shortening the indirect waiting time and increasing the probability of healthcare access. In other words, a welldesigned appointment system should increase the utilization of scarce resources and reduce the total waiting time ( indirect and potential waiting time) of patients [2].

|  |
| --- |
| 1545-5955 © 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications\_standards/publications/rights/index.html for more information. |

In terms of scheduling horizon, an appointment system can be divided into two parts: intraday planning and interday planning, as shown in Fig. 1. In this paper, we focus on interday planning, which takes place before appointment day, rather than intraday planning (scheduling that takes place after the patient arrives at the hospital). Patient waiting time is a critical factor in evaluating the effectiveness of an appointment system, and this parameter can be classified into direct and indirect waiting time depending on the planning horizon [3]. In this paper, we concentrate on increasing patient access probability and shortening patient waiting time during interday scheduling.

Previous studies [2] on appointment rules have focused on designing the length of block service time (or hospital session, which refers to the time spent in updating the appointment systems available slots) and block capacity (number of slots in each hospital session). We include the booking window length in the appointment rule, because it is a critical factor that influences the patients waiting time and access opportunity during interday planning. The appointment rule in this paper, therefore, includes 3-D variables, namely, booking window length, block capacity, and block service time.

Motivated by the current situation of the difficulty in obtaining timely access to hospitals, we model and evaluate the effect of appointment rules on the appointment system, particularly on interday planning. We also consider patient cancellation behavior, which is a critical factor for interday appointment systems [1], [3], [4].

To model the interday appointment system in conformity with practical situations, we establish a renewal process that consists of two parts. The first part uses the *M/D*[*b*]*/*1*/N* bulk service queue to model the service process for patients who have already obtained an appointment. The second part uses the birth and fixed duration death process to describe the state transfer of potential patients who want to gain access to the appointment system. Compared with the previous *M/D/*1*/N* individual service queue [1], [5], [6], our *M/D*[*b*]*/*1*/N* structure is more appropriate for describing the interday planning system, because it considers indirect waiting time and potential waiting time.

An associated embedded Markov chain is introduced to analyze this renewal process. The analytical solution of the steadystate distribution is derived on the basis of the probability generation function approach and the roots-based procedure. With the steady-state distribution, we conduct a numerical study to examine the effect of each appointment rule parameter. Then, we provide an actual case study of three typical departments of PUTH to demonstrate the effectiveness of optimal appointment rules derived from our model.

The remainder of this paper is structured as follows. Section II provides a literature review. Section III describes our problem and stochastic model. In Section IV, we analytically derive the system steady-state distribution. Section V presents numerical experiments and performance measures. In Section VI, we conduct a case study based on an actual case from three typical departments of PUTH. Section VII presents the conclusions and future plans. All proofs are provided in the Appendix.

II. LITERATURE REVIEW

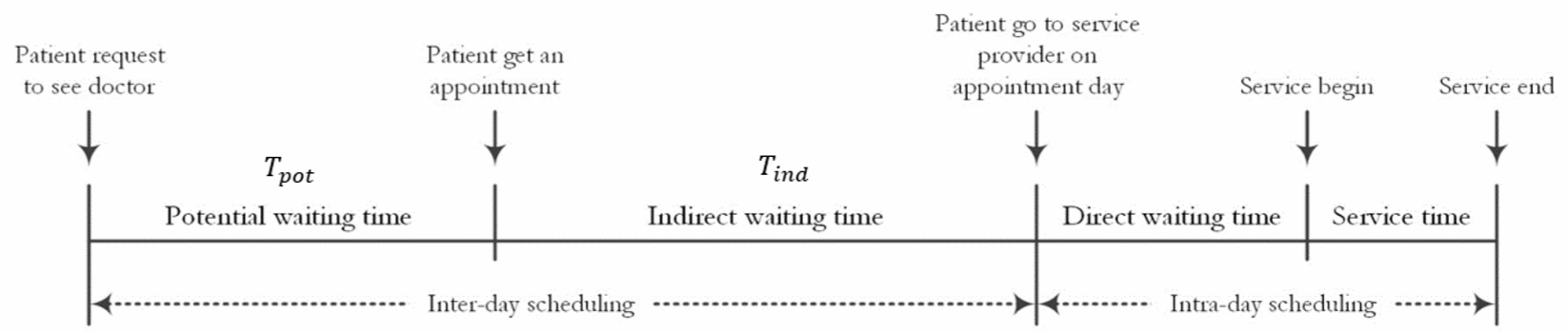
Numerous studies have been conducted to design and evaluate appointment systems, and several of them [2], [4], [7] have provided a comprehensive review. Existing studies can be grouped into intraday and interday planning depending on the scheduling horizon.

Studies on interday planning also focused on improving appointment systems by changing appointment rules, which is more challenging than that for intraday planning problems [4]. This paper partly belongs to this category. However, we pay more attention to a novel patient potential waiting scheme. Existing studies have investigated the patient indirect waiting time to improve patients’ timely access. Reference [4] pointed out that the challenges of modeling patients’ indirect waiting time are the complexity of infinity-horizon planning and matching different patients’ preferences. Researchers thus limited their models to relevant assumptions to obtain tractable analytic solutions. Several researchers [5], [8] regarded scheduled appointments as a single-server queue. In this manner, the infinity-horizon issue turns into an advantage on problem solving in queuing theory. The second purpose is to develop an optimal schedule policy associated with the development of system accessing rules. Models derived from the Markov decision process are introduced to decide whether to accept a patients request and which appointment day to book [9], [10]. Another purpose is to explore the tradeoff between the advanced and same-day appointments. In other words, remaining slots for the same-day appointments are reserved before making an advanced appointment. Reference [11] attempted to allocate the same-day potential patient to present-day slots, and the optimal patient sequence indicated that an advanced appointment must be in a single block at the beginning or in the early middle part of the schedule. The last purpose is to compare and evaluate different appointment systems. Reference [12] claimed that an open-access schedule significantly outperforms a traditional schedule on the basis of the weighted average objective function. However, [13] pointed out that a short booking window system is more capable of smoothening out demand than the open-access system, and the former can improve resource utilization with reduced costs.

With regard to outpatient interday planning, this paper focuses on the adjustment of the appointment system capacity and rules. The work that is most related to ours is [1], which generally models the appointment system as an *M/D/*1*/N* queue with state-dependent no-shows in the quest for the expected patient backlog and the probability of obtaining the same-day appointment. This paper aims to determine the optimal patient panel size. Another relevant work is [6], which proposed a queuing model and a characterized system reward in the objective function to design the optimal appointment window length. This paper focuses on the system congestion, with the aim of regulating the resource supply according to the demand in terms of three system variables (booking window length, block capacity, and block service time). In contrast to previous studies, we consider potential patients to particularly examine the shortage of slot in relation to the potential indirect waiting time.

We also derive a renewal process that combines an

*M/D*[ *b*]*/*1*/N* bulk service queue and a birth and death process to model the practical potential indirect waiting time situation. Closed-form analysis of the *M/D*[*b*]*/*1*/N* queues steady-state probabilities remains a problem. Reference [14] pioneered

Fig. 1. Appointment system planning horizon and patient waiting time.

the research on an infinity bulk service queue on the basis of generating a function approach that estimates the range of average waiting time and queue length without providing a general expression. Most studies have attempted to establish the general expression of these performance measures under different service distributions [15]–[17]. Reference [18] developed a root estimation method to approximate the queue length distribution. Following this method, [19] and [20] performed the computational estimation of the steady-state probability in the *M/D*[*b*]*/*1*/N* problem and proposed the closed form of the steady-state waiting time distribution. For the finite bulk service queue, [21] developed the embedded Markov chain of *M/D*[*b*]*/*1*/N* queue and converted this memory-less process into a discrete Markov chain, but the analytical solution can be obtained only in the *M/D/*1*/N* queue [22], which is a special case of *M/D*[*b*]*/*1*/N*. Similarly, obtaining the closed form of our renewal process steady-state distribution is difficult, because the nonuniform system batch transfer makes the transition matrix intractable. We therefore start by modeling our renewal process in a transient analysis, introducing the embedded Markov chain, and using a roots-based generation function approach to obtain the analytical solution of the steady-state distribution, which is the contribution of this paper in terms of technique development.

III. MODEL FORMULATION AND PRELIMINARIES

*A. Problem Description*

We consider the interday appointment system in Fig. 2 and assume that patient requests follow a Poisson process with arrival rate *λ*. This assumption has been validated in [23]. This system is defined by the appointment rule *A* := *(T,n,l)*, which has a finite slot capacity *N* during a finite booking window *T*. Booking window *T* is divided into *k* blocks, each with equal length of periods *l* = *T/k*. In each block, if *n* = *N/k*, an equal number of slots exist. This batch of *n* slots can be regarded as the capacity of a hospital session or an appointment block. For example, *A*1 = *(T*1 = 10*,n*1 = 20*,l*1 = 1*)* and *A*2 = *(T*2 = 10*,n*2 = 1*,l*2 = 0*.*05*)* are typical systems that possess the same total capacity *N*1 = *N*2 = 200 during the same booking window *T*1 = *T*2 = 10. The former indicates a daily updating single-block system with *k*1 = 10, whereas the latter is a real-time updating individualslot system with *k*2 = 200.

Following earlier works [1], [5], we assume that an appointment patient will be served in a first-come, first-served sequence. This assumption means that the system schedules

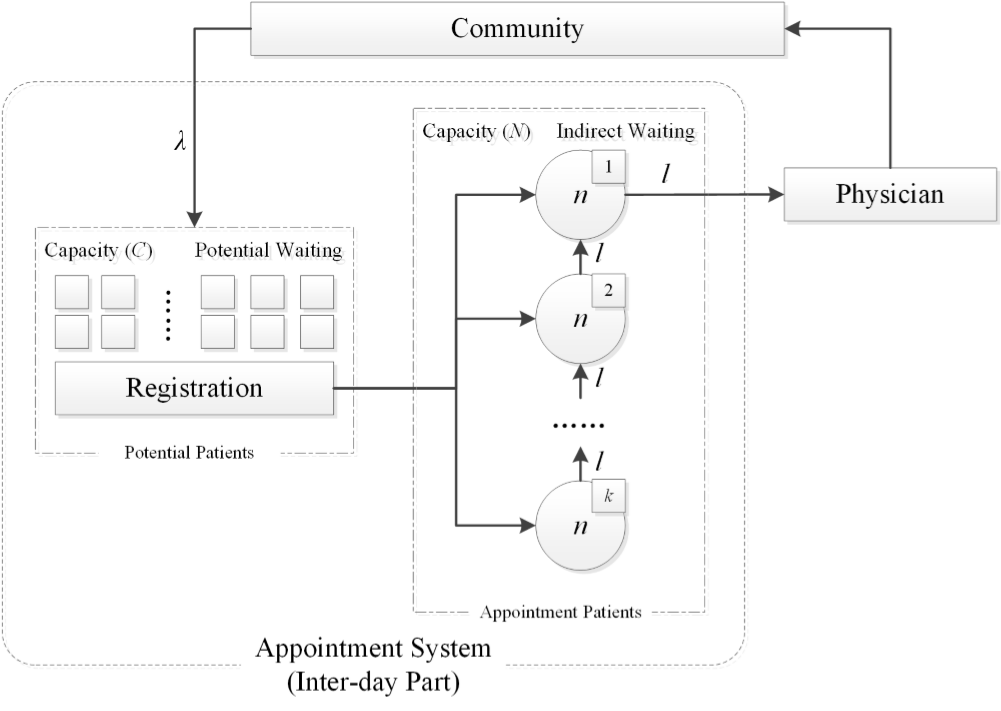


Fig. 2. Schematic representation of the interday appointment system.

a new appointment request to the earliest available block. This assumption is appropriate under situations of high arrival rates and limited appointment slot resources. Earliest available slots are always patients’ first choice. Moreover, scheduling to the earliest slots is reasonable for systems with long indirect waiting time, because timely access is also an important patient requirement.

Congestion will occur if *N* slots are fully booked, and additional requests will fail to gain access, because the system has a finite capacity *N*. This paper particularly considers these potential patients and aims to improve their access performance by designing enhanced appointment rules. We divide the system into two parts on the basis of the status of whether the patients have gained access to the appointment or not. First, patients who have successfully made an appointment become the doctors’ future backlog. Cancellation by these patients with appointment occurs during interday appointment scheduling. We denote *c,*0 ≤ *c* ≤ 1 as the cancellation rate. This paper does not consider the patient no-show, because noshow is a situation that only happens in intraday planning. Idle slots emerging from the patient cancellation are rescheduled by new incoming requests. Patients with appointment in the earliest batch (e.g., patients in node 1, Fig. 2) who are going to see a doctor in an immediate block depart together in the update period. At the same time, the system automatically shifts other blocks up to forward ones and therefore releases a batch of new slots. This process can be formulated by a single-server queuing system with bulk service. Second, for potential patients who have failed to gain access, their request continues after leaving without obtaining an appointment. To study these potential patients, we design a *fictitious pool* to reserve them. Infinite space is available for this fictitious pool to accommodate all potential requests that appear during this block. This fictitious pool is only for one blocks potential patients and is dismissed at the end of this block. It is a Poisson birth with a deterministic time death process. We use it to count the number of patients that failed to gain access in each block.

In summary, the following four actions transpire simultaneously in the block update period: first batch leaving, and other batch shifting up, new batch releasing, and potential patient pool clearing up.

*B. Model Description*

Considering patients with appointment and potential patients, a renewal process is defined. The state space under appointment rule *A* is *SA* = *iA*, where *iA* ∈ {0*,*1*,*2*,*3*,...*} is the total number of patients in the appointment system. We let *XA(t)* be the number of patients in the system at time *t* under appointment rule *A*. The stochastic process {*XA(t)*|*t >* 0} is a renewal process with state space *SA*. We let *πiA (t)* denote the probability that *iA* patients are in the system at time *t*. We use *A(t)* = *πiA (t)* to represent the probability distribution of the state at time *t*. Fig. 3 shows the transition diagram of the appointment system.

Arrival rate *λ* follows the Poisson assumption and system state *i* goes to *i* + 1 after each individual arrival. Here, we denote *ρ* = *λl/n* as the traffic coefficient, which indicates the relationship between demand and resources. If *ρ >* 1, it means that in each block, the number of arrivals is more than the number of slots, and vice versa. We let *αi* denote the probability that *i* patients arrive during a batch service time *l*.

We, then, have

*αi* = *(λl*!*)i e*− *λl.* (1)

*i*

According to the batch service design, the departure process has the following state transition:

⎧0; *i* ∈ {0*,*1*,...,n* − 1} ⎪

*i* −→ ⎨*i* − *n*; *i* ∈ {*n,n* + 1*..., N*}

⎪⎩*N* − *n*; *i* ∈ {*N* + 1*, N* + 2*,...*}*.*

The departure process is not time-memoryless, because the system service time of a batch is constant, which means that we have to specifically define their rates. We let *DiA (t,t*+*t)* be the probabilities that a batch departure occurs between time *t* and *t* + *t* when the system has *i,i* ∈ {0*,*1*,...*} customers in queue under appointment rule *A*. When *i* = 0, no patient departs. Moreover, the corresponding departure rates are

*diA(t)* = lim *.* (2) → *t*

To obtain the steady-state probability *iA* analytically, we introduce an associated embedded Markov chain in

Section IV.

IV. MODEL ANALYSIS

*A. Embedded Markov Chain*

The appointment patient set and the potential patient set are updated at the end of a block, just at the time when departure occurs. We set up an embedded Markov chain with the same update epoch as our appointment system. We let *t r* be the time of the *r*th transition epoch, which means *t r* = *(t* − 1*)l, r* ∈ {1*,*2*,...*} . In this manner, the embedded Markov chain is {*X A(tr)*}*, r* ∈ {1*,*2*,...*}, where *XA(tr)* denotes the number of patients in the system just before a departure occurred at time *t r* under appointment rule *A*, which has the same state space *S A* = {*iA*} as our original renewal process. The one-step transition probability matrix of the embedded Markov chain

{*XA(tr)*} with *p(i,j)A* is given by *p (i,j)A* = *αj*; *i* ∈ {0*,*1*,...,n*}*, j* ∈ {0*,*1*,...*} *p(i,j)A* = *αj* −*(*min[*i,N*]−*n)*; *i* ∈ {*n* + 1*,...*}

*j* ∈ {min[*i, N*] − *n,...*}

*p(i,j)A* = 0; otherwise*.* (3)

We denote {*qiA*} as stationary probabilities. Two processes are connected by the relation

*d iA*

*qiA* = ∞*j*=1 *diA , i* ∈ {0*,*1*,...*}*.* (4)

*Corollary 1:* The embedded Markov chain is irreducible, aperiodic, and positive recurrent. Stationary probabilities of this process exist and have the following property:

|  |  |  |
| --- | --- | --- |
| lim *qiA* = 0  *i*→∞  ∞  *qiA* = *p(j,i)AqjA,*  *j*=0  ∞  *qiA* = 1*.* | *i* ∈ {0*,*1*,...*} | (5)  (6)  (7) |

*i*=0

According to Corollary 1, we can get the stationary probability *qiA* , which is the unique nonnegative solution of (6) and (7). Stationary probabilities *qiA* are the departure distribution at an update epoch, and *qiA* represents the probability of *i,i* ∈ {0*,*1*,...*} patients in the system at block ending time.

We will also refer to *q iA* as departure distribution in the following. The size of our state space is infinity, which indicates that the stationary distribution cannot be directly calculated by simultaneously solving the inverse matrix of (6) and (7). A probability generating function approach is expanded to derive the recursion of stationary distribution in Section IV-B.

*B. Generating Function Approach*

We denote *PA* *qiA zi* (*z* complex with |*z*| ≤ 1) as the generating function of departure distribution {*qiA* }, which can help to develop the recursion between *qiA* values. The basic process of the generating function approach is to first find the closed-form expression for *PA(z)* using (6).

*Corollary 2:* The probability generating function of the departure distribution is

*n*−1

−

*z*

*n*

*)*

−

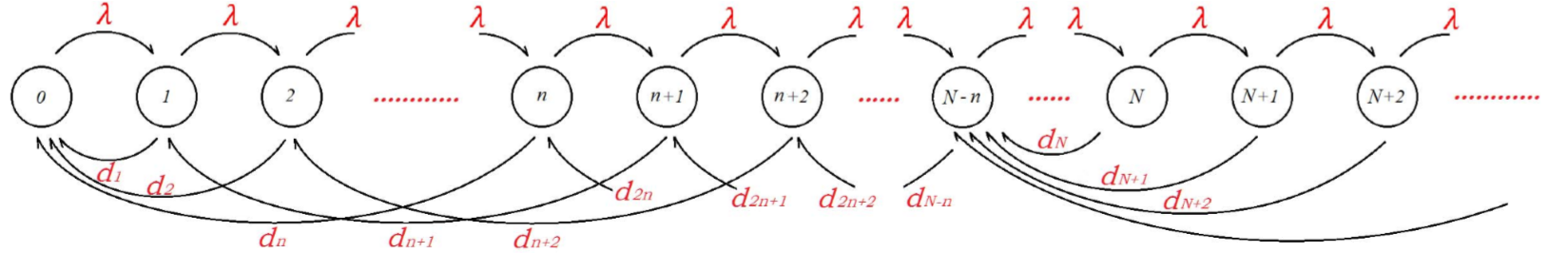
∞

*i*

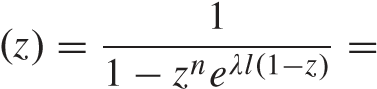


*P A(z)* = *i*=0 *qiA (zi* − *neλl(*1−=*zN) qiA (zN* − *zi).* (8)

1 *z*

Fig. 3. Transition diagram of the interday appointment system.

Then, we expand *PA(z)* in a power series, and the departure distribution {*qiA*} can be obtained as the coefficients in this power series. The numerator of (8) has already satisfied the requirement, and an application of identical transformation will rewrite the numerator as a power series of *zi*. However, the denominator of (8) needs to be converted into a power series form following the formulation:

 ∞ *i*

*BAbiA z ,* |*z*| ≤ 1 (9)

*i*=0

and proposition 1 provides a recursion of stationary distribution {*qiA*} with the closed-form expression for coefficient *biA*.

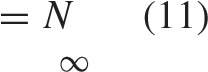
*Proposition 1:* The stationary distribution {*qiA* } of the embedded Markov chain is given by

*n*−1 *n*−1

*qiA* = *qjAb(i*−*j)A* − *b(i*−*n)A qjA, i* ∈ {*n,..., N* − 1} *j*=0 *j*=0

(10) *n*−1 *n*−1 ∞

*qiA* = *qjAb(i*−*j)A* − *b(i*−*n)A qjA* − *b*0*A*  *qjA j*=0 *j*=0 *j*=*N*+1

*i* *n*−1 *n*−1

*qiA* = *qjAb(i*−*j)A* − *b(i*−*n)A qjA* − *b(i*−*N)A*  *qjA j*=0 *j*=0 *j*=*N*+1

∞

+ *qjAb(i*−*j)A, i* ∈ {*N* + 1*,...*} (12)

*j*=*N*+1

where *b*0*A* = 1 and for *i* ≥ 1

*i*  *( λlj)i*−*jn*

*b iA* = *n* − − ! *eλlj.* (13)

*(i jn) j*=0

The basic generators in formulas (10)–(12) are the probabilities {*q*0*A,q*1*A,...,q(n*−1*)A*} and {*qNA,q(N*+1*)A,...*}, where the former vector is the initial probabilities and the latter is the tail probabilities. To derive the analytical solution of *qiA*, the expression of the initial and tail probabilities must be determined at first. In Corollary 3, we rewrite (11) and eliminate the tail probabilities. We then implement the distribution normalization condition (7) to demonstrate the relationship of initial probabilities.

*Corollary 3:* The sum of tail probabilities is given by

∞ *n*−1 *n*−1

*qiA* = *qjAb(N*−*j)A* − *b(N*−*n)A qjA.* (14)

*i*=*N j*=0 *j*=0

The initial probabilities {*q*0*A,q*1*A,...,q(n*−1*)A*} obey the following normalization equation:

*n*−1 *n*−1

*qiA b(N*−*j)A* = 1*.* (15) *i*=0 *j*=*i*

From now on, each tail probability *q iA*, *i* ≥ *N* can also be expressed by these initial probabilities. We denote *βijA* as the coefficient of initial probabilities *qjA*, 0 ≤ *j < n*, which means *n*−1 *h*=0

|  |  |  |
| --- | --- | --- |
| where | *qiA* = *βijAqjA, i* ∈ {*N, N* + 1*,...*}  *j*=0 | (16) |
|  | *N*−*n βijA* = *αi* + *αi*−*h*[*b(h*+*n*−*j)A* − *bhA*]*.* | (17) |

The initial probabilities are now the only generators of the interday appointment system. The practical interpretation of

isCorollarythe average3 is thatprobabilitythe sumwhenof tailnoprobabilitieswaiting space ∞*i*=is*N* left *qiA*

at the departure epoch in an *M/D*[*b*]*/*1*/N* queue system. In Section IV-C, a roots-based calculator is developed to derive the steady-state distribution.

*C. Roots-Based Calculator*

Following [20], which applied the roots-based method on the infinity bulk service process, we extend it on our combination model to derive the initial probability. After the normalization, the probability generating function (8) satisfies *PA(*1*)* = 1. However, the denominator of (8) is equal to 0 when *z* = 1.

According to Rouché’s theorem [18], we can show that the equation 1 − *zneλl(*1−*z)* = 0 has *n* distinct roots *z*0*,z*1*,...,z(n* −1*)*, which satisfies |*z*| = 1 and *z*0 = 1. The numerator of (8) must have the same roots, because *PA(z)* is also convergent at the same range. We then rewrite the probability generating function (8) with roots as

*z j)*

[

*R*

−

*Q*

*(*

*z*

*)*

]

*(*

*z*

−

1

*)*

*n*

−

1

*j*

1

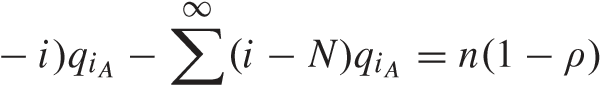
*(*

*z*

−



*P A(z)* = − *neλl(*1−*z)* = *,* |*z*| ≤ 1 (18) 1 *z* where *R* and *Q(z)* are the undetermined coefficient and polynomial of the numerator after reduce the denominator. Applying L’Hospital’s rule both in (8) and (18) and letting *z* = 1 yield

*n*−1 

*(n*(19)

*i*=0 *i*=*N*

and the undetermined coefficient *R* can easily be written as

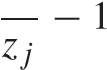
*R* = *ni*=*n*−*j*−= 01 11*((i*1−−*nz)qj)iA .* (20)

Equation (19) has a probabilistic interpretation. The first term on the left-hand side provides the average number of underutilized slots of the first block, and the second term is an average slot shortage. The difference between these values is equivalent to the mismatch of supply and demand. By replacing *R* of (18) with the expression in (20), initial probabilities {*q*0*A,q*1*A,...,q(n*−1*)A*} can then be calculated by the inversion of generating function *PA(z)*.

*Proposition 2:* The initial probabilities obey the following recurrence formulas:

*qiA* = *γiA q*0*A, i* ∈ {1*,...,n* − 1} (21)

where *γiA* is the coefficient of *zi* in the product

*n*−1 *γ A z i* = *n* −1 *z .* (22)

*i*

*i*=0 *j*=0

Proposition 2 builds the relationship of initial probabilities, and *q*0*A* is now the only generator of our embedded Markov chain. By solving the simultaneous equations, the roots-based recurrence (21) and the normalization equation (15) of initial probabilities, the value of *q*0*A* can be calculated

⎛*n*−1 *n*−1 ⎞−1 *q*0*A* = ⎝*γiAb(N*−*j)A*⎠ *.* (23) *i*=0 *j*=0

In summary, the procedure for computing stationary distribution *qiA* is as follows.

Step 1: Compute− *neλl(*1the−*z)*roots= 0. *z*0*,z*1*,...,zn*−1 of the equation

1 *z*

Step 2: Calculate the initial probabilities with (21)–(23).

Step 3: Calculate the middle probabilities with the first equation of (10).

Step 4: Calculate the tail probabilities with (16) and (17).

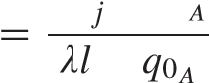
Here, the three parts of the stationary distribution *qiA* are defined as follows.

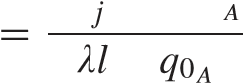
1. *Initial Probabilities:* {*q*0*A,q*1*A,...,q(n*−1*)A*}.
2. *Middle Probabilities:* {*qnA,q(n*+1*)A,...,q(N*−1*)A*}. 3) *Tail Probabilities:* {*qNA,q(N*+1*)A,...*}.

These four steps develop a procedure that can help us analytically derive the departure distribution. With the departure distribution *qiA* , system steady-state distribution *A* can be directly derived on the basis of the following proposition.

*Proposition 3:* The steady-state probabilities *πiA* and departure distribution *qiA* are related by

*n* 0 *qj*

*πiA* = *, i* = 0 (24a) +

*i*+*n i*+1 *qj*

*πiA* = + *, i* ∈ {1*,*2*,..., N* − *n*} (24b)

∞ *q*

*πiA.* (24c)

*D. Patient’s Cancellation Behavior*

Finally, considering the cancellation and rescheduling process, only appointment patients are capable of cancellation under probability *c*. In this way, patients who have long indirect waiting time potentially have a high possibility of canceling their appointments. Potential patients can be scheduled into the system if idle slots are canceled by appointment patients. This cancellation process is a binomial distribution *(i,c)* in which the number of cancellation patients takes a value equal to *m*, and we let this probability to be *B*[*i,c,m*], that is

[ ] = *i*  *m (*1 − *c)(i*−*m).* (25)

*B i,c,m c m*

Patient cancellation and slot rescheduling will occur in one block, and the system steady-state distribution {*πiA*} must be consequently changed. In the long run, we only consider the system state *S A* = {*i*}*,i* ∈ {0*,*1*,...*} and compare the

stationary probability before and after cancellation. We let *π*ˆ*iA* be the adjusted system stationary probability when the system has *i* patients after bringing in cancellation and rescheduling.

We have

∞

*π*ˆ*iA* = *B*[*j,c*ˆ*, j* − *i*]*πjA, i* ∈ {0*,*1*,...*} (26)

*j*=*i*

where *c*ˆ = 1 − *(*1 − *c) l* is the patient cancellation rate for each block. *π*ˆ*iA* only comes from original probabilities *πiA* whose states are equal to or greater than this adjusted probability. The reason is mainly because the largest cancellation population should not exceed the total number of patients in the system. When *c* = 0, which means no cancellation exists, then *π*ˆ*iA* = *πiA* . Our cancellation and rescheduling process only concern the system state in the long run but do not regard the patient position changing in the sequence. Adjusted steady-state distribution {*π*ˆ*iA* } can therefore replace the original distribution {*πiA* }.

V. PERFORMANCE EVALUATION

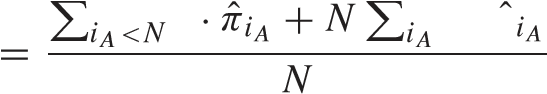
With the steady-state distribution, we can now evaluate the appointment system under different appointment rules. As mentioned earlier, a well-designed appointment system should balance resource utilization, patient access probability, and indirect waiting time. We thus define three main measures in the following to evaluate our appointment system:

*P*LWBS*(λ,c, A)* = *π*ˆ*iA* (27)

*i A*≥*N*

*L* app *U*slots*(λ,c, A)* = 

*N*

*i* ≥*N π* (28)

*T*total*(λ,c, A)* = *T*ind*(λ,c, A)* + *T*pot*(λ,c, A)*

= *l* *i*  *l* *N* · *π*ˆ*iA*

*n iA<NiA*≥*N λl*

+ − *.* (29)

*N L*app

TABLE I

DEVIATION RATE OF SYSTEM PERFORMANCE MEASURES

First, *P*LWBS is the proportion of time that congestion occurred in our interday appointment system. It represents the probability that a patient will fail to access and leave without being seen (LWBS) based on a long-run condition. This index can help hospital policy makers understand system congestion level. Second, *U*slots is the long-run slot utilization, which is a ratio between the average number of appointment patients and total slots. This value shows the occupancy rate of slots for a single server that can help understand if the design of the total slot capacity is reasonable. Finally, *T*total is a summation of the patient indirect waiting time *T*ind and the potential waiting time *T*pot. The latter is the forecasted average time that a patient may obtain an appointment. When *T*pot ≤ 1, the system has superfluous slots and patients directly gain access. Otherwise, patients keep attempting to seek an appointment in the future block if they insist on meeting this doctor.

On the basis of the above-described performance measures, the system can be evaluated by the given patients arrival rate *λ*, cancellation probability *c*, and appointment rule *A*, which includes system booking window length *T*, block capacity *n*, and block service time *l*. Patients’ arrival rate and cancellation probability are exogenous factors that can be determined in advance. The appointment rule is the decision variable.

*A. Comparison With Simulation Results*

To verify our model, we establish a discrete-time event simulation model and compare the simulation results with the analytical solution in terms of three system performance measures.

We consider two types of events with a different randomness. Patient arrival and cancellation randomly occur during an update period. A block service deterministically occurs in the update period. A total of 800 different combinations of system input parameters *C(λ,c, A)* are available, including patient arrival rate *λ* ∈ {2*,*4*,...,*20}, probability of the patient cancellation *c* ∈ {0%*,*10%}, length of booking window *T* ∈ {7 days*,*15 days}, block capacity *n* ∈ {2*,*4*,...,*20}, and duration of block service time *l* ∈ {1 day*,*0*.*5 days}. The permutations and combinations of these input parameters have 800 different system input cases. The simulation is carried out 20250 days for each case, and the first 250 days are the warmup time.

The average deviation rate *Δψ* of the three performance measures is shown in Table I, and

*λ c A* *fψ* *(λ,cf,ψA()λ,*−*cf,ψA() λ,c,A)*



*Δψ* =

800

where *fψ(λ,c, A)* and *fψ* *(λ,c, A)* are the simulation results and analytical results, respectively, for the performance

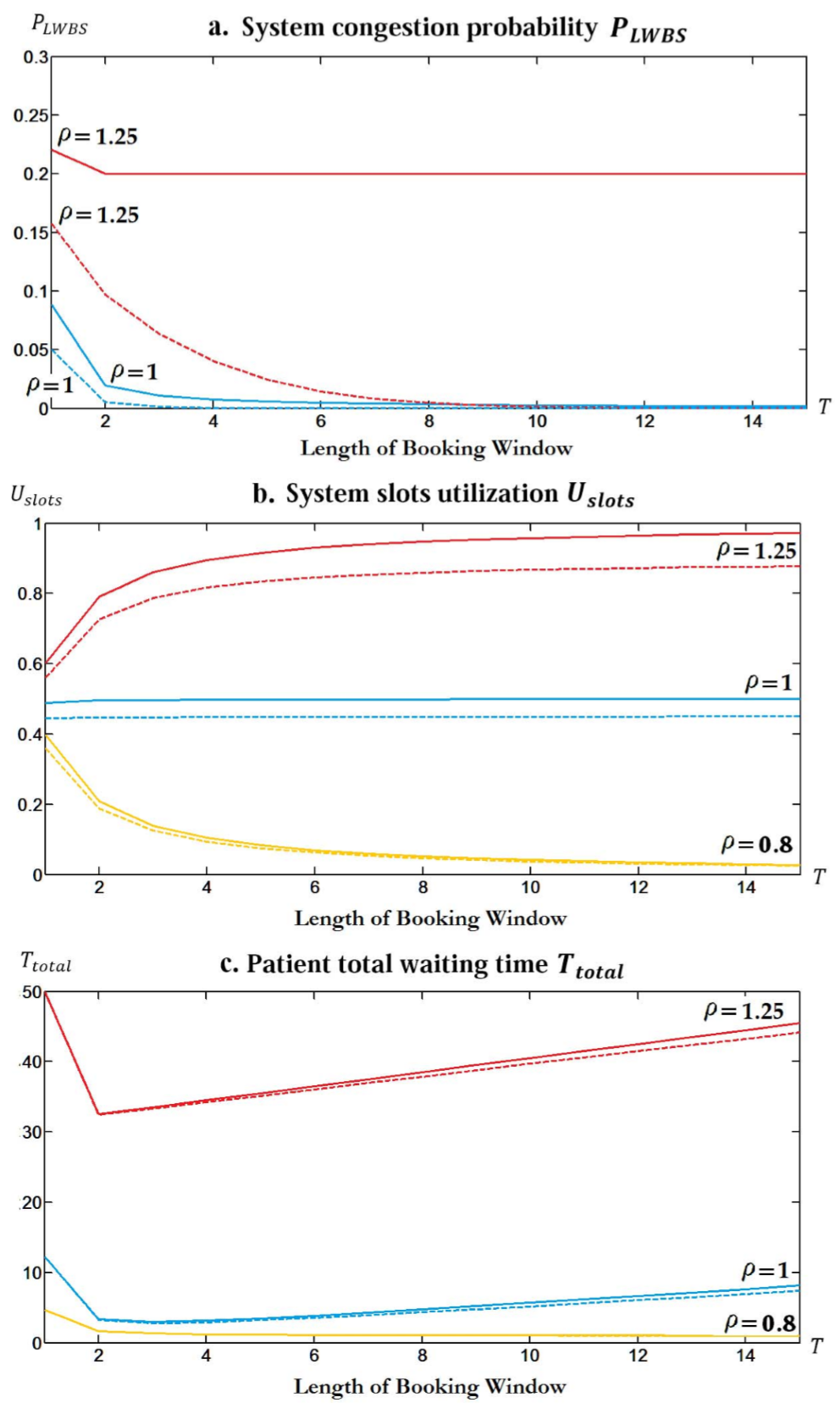


Fig. 4. Performance measures under different appointment rules *(T,(*20*/ρ),*1*)(λ* = 20*)*. Solid line for *c* = 0 and dashed line for *c* = 10 %.

(a) System congestion probability *P*LWBS. (b) System slots utilization *U*slots. (c) Patient total waiting time *T* total.

measure *ψ*. Table I shows that the deviation rate for all performance measures is below 5%, and it is only 0.83 % for patient congestion probability. This low deviation rate indicates that our above-mentioned analysis is quite accurate.

*B. Sensitivity Analysis*

In this section, we illustrate the interaction between the parameters of appointment rules and system performances.

*1) Length of Booking Window:* Fig. 4 shows the three performance measures under varying booking window lengths as follows: congestion probability *P*LWBS, resources utilization *U*slots, and patient total waiting time *T*total, with different system traffic coefficient *ρ* values.

For the patient congestion probability in Fig. 4(a), an interesting feature of those lines is that they all level off after a short booking window, especially in the case that the supply cannot satisfy the demand (*ρ >* 1). The dashed lines suggest that system congestion probability decreases when the cancellation rate reaches 10%. This result shows that lengthening the booking window is useful to reduce congestion probability, especially when the cancellation exists.

For resource utilization *U*slots in Fig. 4(b), the effect of patient cancellation is limited. The changing directions of *U*slots are different under varying *ρ* with a long booking window. Separated by the condition of *ρ* = 1, system resource utilization increases with the lengthening of booking window when *ρ >* 1, while it goes against when *ρ <* 1. According to this circumstance, it is meaningful to design a long booking window for an insufficient supply system but useless for an over supplied system.

This is mainly because we use the *M/D*[*b*]*/*1*/N* bulk service queue to model the service process of patients who have already got an appointment, as shown in Fig. 2. Lengthening the booking window (*T*) results in more slots (*N*), which reduces the impact of the blocking and stochastic effect. When *ρ >* 1, more slots result in a pooling of the uncertainty that the system is not full, and therefore increases the utilization of the slot. While, when *ρ <* 1, the average queue length must be less than a constant, which can be calculated for the queue without buffer limitation. Hence, *U*slots is decreasing in *N*, when *ρ <* 1. As for the case *ρ* = 1, since the rate in and rate out of each state are equal, in the steady-state distribution, the probability of each state is equal. Hence, the utilization of the slots is always 0.5. In addition, since the number of patients’ cancellation is modeled as a binomial distribution, the higher the utilization of slots, the larger the expected number of cancellations. This is why lengthening the booking window is useful to reduce congestion probability, especially when the cancellation exists.

For the patient total waiting time *T*total in Fig. 4(c), a short booking window system exhibits optimal performance when considering the patients timely access by minimizing *T*total. Analysis of the components of *T*total shows that the patient indirect waiting time increases linearly with the lengthening of the booking window. This result inherently explains why a long booking window system is an inefficient means to deal with patient timely access.

*2) Block Service Time:* We consider the change in block service time *l* under fixed booking window length *T*. The Fig. 5 displays three performance measures under varying number of daily blocks 1*/l* and different system traffic coefficient *ρ* values.

Fig. 5(a) shows that the probability of patients failing to access is unaffected by varying the number of daily blocks when no cancellation occurs in the system. However, if cancellation occurs, then a sparse appointment block can allow more patients to gain access. For the resource utilization *U*slots in Fig. 5(b), the changes are negligible. Fig. 5(c) shows that block service time exerts an insignificant effect on appointment and patients waiting time. These pieces of evidence suggest that for providers with insufficient supply, a sparse appointment block may be a good choice. Sparse segmentation of daily blocks can reduce the system congestion probability and exerts a negligible effect on resource utilization and patients waiting time.

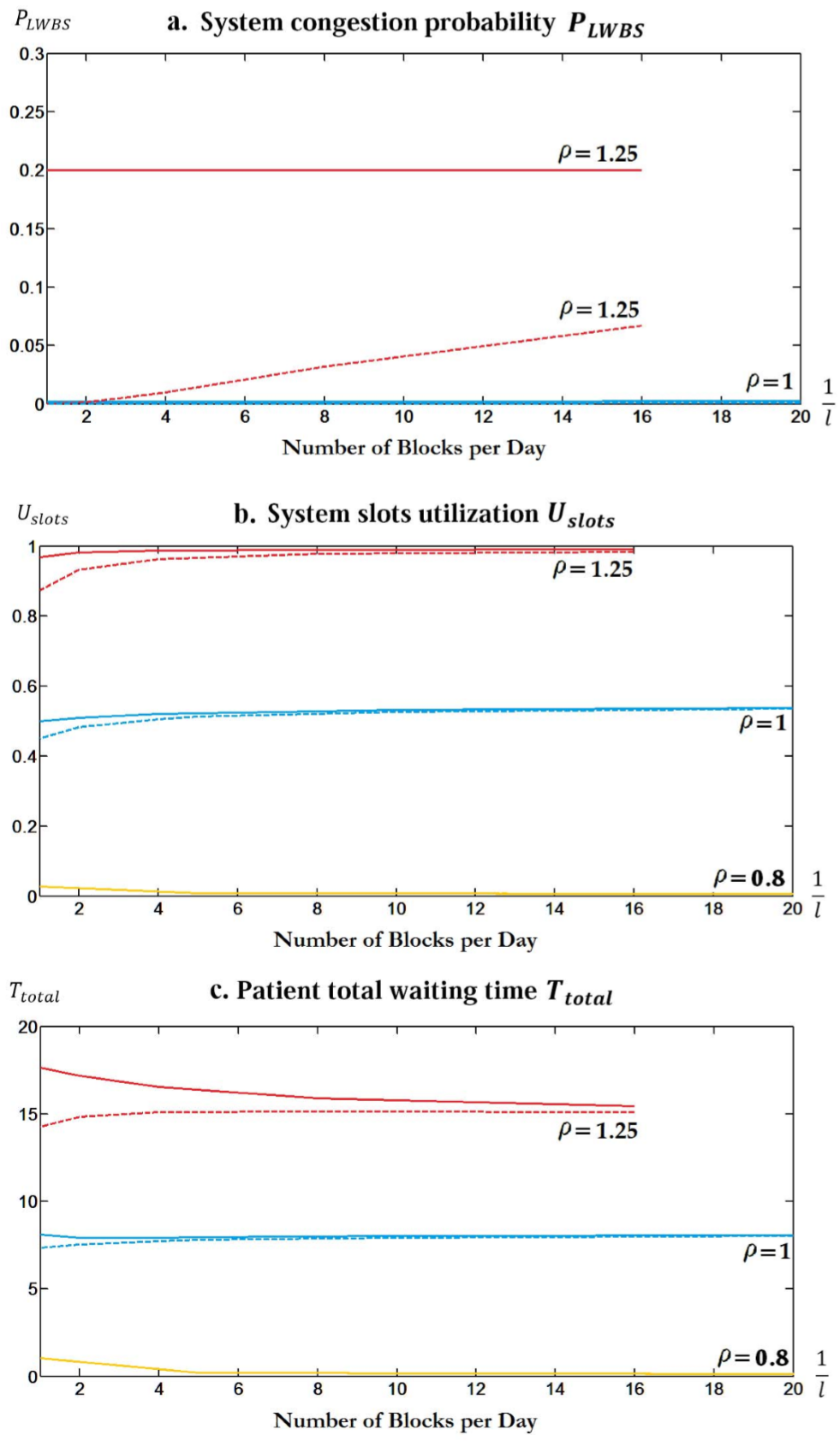


Fig. 5. *P*LWBS under different appointment rules *(T* = 15*,(*20 *l/ρ),l)(λ* = 20*, c* = 0*)*. Solid line for *c* = 0 and dashed line for *c* = 10%. (a) System congestion probability *P* LWBS. (b) System slots utilization *U*slots. (c) Patient total waiting time *T*total.

VI. CASE STUDY

To demonstrate how our model can be used to improve the appointment system of hospitals, we provide an actual case study on three typical departments of PUTH based on our model.

The purpose of designing an efficient appointment system is to minimize an average long-run cost, including service provider operating cost, appointment patients waiting cost, and potential patients opportunity cost, by changing appointment rule *A*. We thus define *C(λ,c, A)* as the long term system cost, which is a weight summation of the costs

*n*

*C(λ,c, A)C*1  + *C*2*T*total*(λ,c, A)* + *C*3*λP*LWBS*(λ,c, A).*

=

*l*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| TABLE II  VALUES OF THREE DEPARTMENTS’ ARRIVAL RATES, CURRENT APPOINTMENT RULES, TRAFFIC COEFFICIENT *ρ*= *λl/n*, AND COST COEFFICIENTS   |  |  |  | | --- | --- | --- | |  |  | | | |  |  | | |  | | | |  |  |  | | |  |  | |   TABLE III  OPTIMAL SYSTEM DESIGNS IN THREE TYPICAL DEPARTMENTS MINIMIZING AVERAGE WAITING   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  | | |  | | | |  |  |  |  |  |  | |  |  |  |  |  |  | |  |  |  |  |  |  | |  |  |  |  |  |  | |

The first term is the daily average operating costs, where *C*1 represents the unit cost for providing a single slot and *n/l* provides the total capacity per day. The second term is used to penalize the hospital for the appointment patients’ waiting. This term is a conservative estimation of patient waiting as explained previously, and *C*2 is designed as the waiting cost per day. The third term represents the long-run average opportunity cost originating from potential patients. Considering that the number of potential patients is *λP*LWBS each day, their opportunity cost is equivalent to the lost revenue of those access-failure patients, and *C*3 is the unit revenue that the hospital can obtain from treating a single patient.

The value of *C*1 and *C*3 can be estimated by the data from the *Annual Chinese health statistics report* and the *Beijing municipal health bureau*. In the implementation of our method, the manager can set the value of *C*2 according to the relative impact of congestion on their operation. For example, for a Chinese hospital, the manager should pay more attention to the service level of specialists’ service and set a higher cost regarding famous specialist.

The design of the interday outpatient appointment system can be simplified and formulated as the optimization problem in the following:

min *C(λ,c, A)*

*A*

subject to *T* ≤ *T*max *n/l* ≤ *s*max

0 ≤ *l* ≤ 1

*T,n* are positive integer*.* (30)

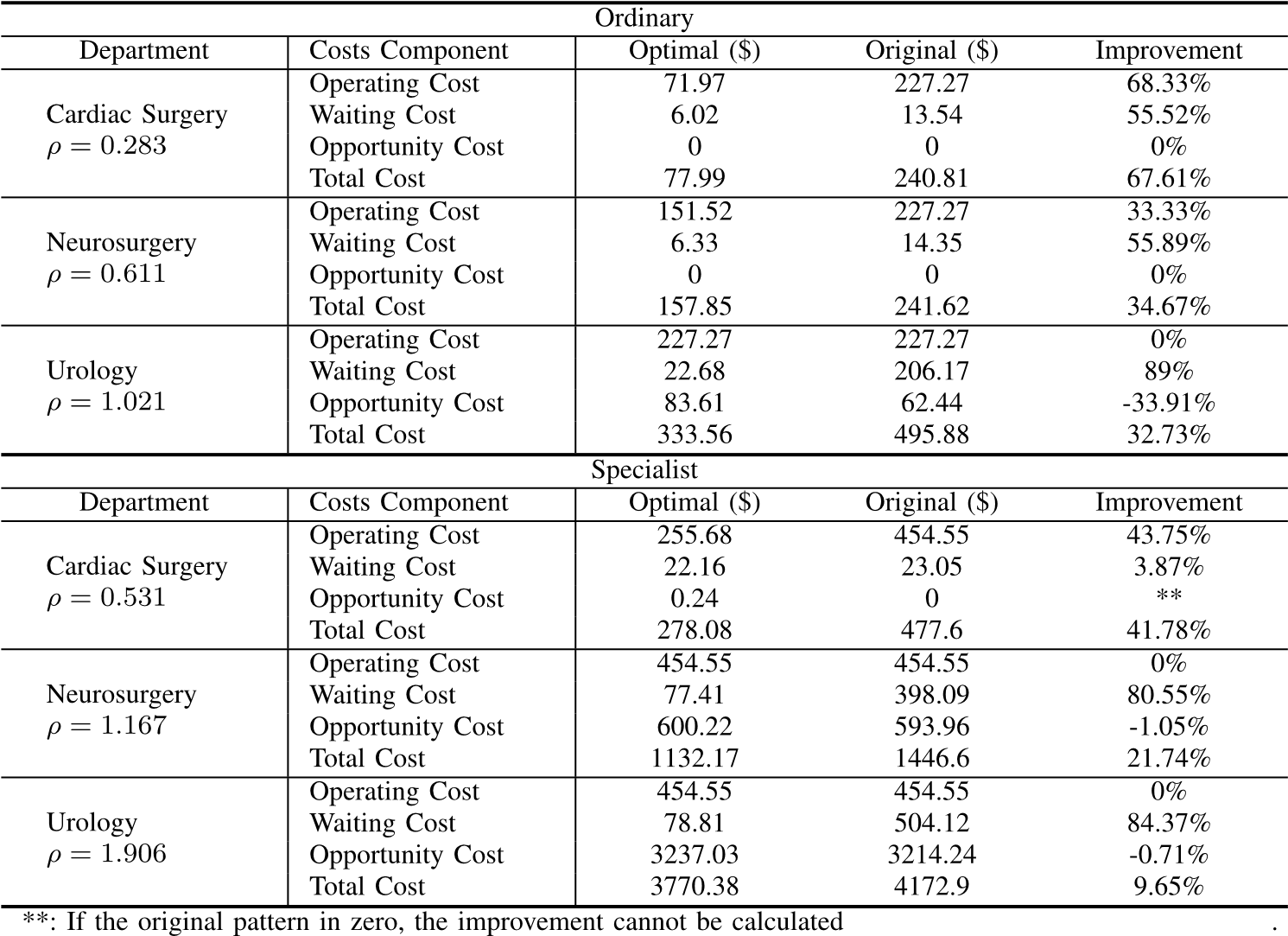
The optimal appointment rule is determined by minimizing the system long-term daily cost function under limited resources. The first two constraints provide the longest booking window length *T*max and the largest daily slot capacity *s*max that the system appointment rule can select. For the third constraint, the duration of one service block cannot be longer than one day, which means at least one block exists per day. With the input of patients arrival rate and cancellation probability, finite feasible appointment rules exist, and the one with the lowest cost will be an optimal choice of the resource-limited appointment system that fits distinct patients demand. On the basis of the programming model in (30), an actual case study is provided, and optimal appointment rules are suggested.

*A. Input Data*

In this section, we collect data of PUTH (a well-known general hospital in China and located in Beijing). In PUTH, registrations of outpatient departments can be mainly classified into two levels: ordinary registration and specialist registration. Specialist registration charges higher and attracts more patients, because its doctors have better reputation. Managers need to decide on the following: the service capacity per day and the booking window length. The former can be represented by daily largest slots *n/l* and *T* with respect to the booking window under appointment rule *A*.

TABLE IV

ORIGINAL AND OPTIMAL COSTS IN THREE DEPARTMENTS



|  |  |
| --- | --- |
| We examine all three typical departments along with the two registration types. A total of six scenarios, which include *ρ <* 1 and *ρ >* 1 cases, are present. By extracting data from the *Annual Chinese health statistics report* and the *Beijing municipal health bureau*, we acquire values of relevant input parameters, including the daily slot limitation, the patient arrival rate, and the maximum booking window in PUTH as Table II shows. The data in Table II indicate that the ratio of the demand (arrival) and capacity (slot limitation) of specialists is higher than that of ordinary doctors in a single department, which is consistent with our assumption.  We solve the optimization problem [as shown in (30)] to minimize the overall system cost. We also consider various cancellation rates, because the relevant data on the cancellation behavior in China are limited. The optimal appointment system configuration represented by the number of slots per block, the time duration per block, and the booking window horizon is obtained.  *1) Optimal Appointment Rule Design:* We first discuss the optimal system designs, which are also called appointment rules. Table III shows the optimal designs for ordinary and specialist doctors, where the different ratios of demand and supply cause diverse required scheduling rules to improve performance. Nevertheless, concerted insights can be refined from Table III despite the diversity of system designs in different scenarios.  • The optimization results confirm our assumption that the ratio of demand and supply (*ρ*) is the crucial factor influencing the system design. The column of the “capacity load” reflects the proportion of assigned capacity | *(n*∗*/l*∗*)/(n/l)* in the optimal solution. The general trends of the optimal capacity load increase as parameter *ρ* increases as if *ρ <* 1. More specifically, the optimal capacity load is always a bit higher than *ρ*, which means that the optimal allocated capacity slightly exceeds the arrival rate. When *ρ* ≥ 1, the changing of *ρ* cannot bring further movement of the capacity load, where the doctor is now working at full capacity.  • As for departments with under loaded doctors (*ρ <* 1) , the optimal size and the duration of blocks are quite small. According to Table III, the single block with one individual in a block is preferred. As for departments with overloaded doctors (*ρ* ≥ 1), the optimal size of the block changes with different cancellation rates in accordance with the trend that the cancellation rate increases, the larger blocks with more than one slot are preferred.  *2) Optimal Value of Cost Function With Improvement:* Cost of the original solution (current state) and the optimal appointment rules is provided in this section. We compare these values in terms of the total cost and its three components. Table IV shows the original and optimal costs, including the labor expense, the LWBS patients cost, the waiting cost, and the total cost as well as a comparison among them. Table IV illustrates that the optimal solution dramatically retrenches the overall cost in all types of departments, which verifies the effectiveness of our optimization model. Further insights obtained from the data shown in Table IV can be expressed as follows.  • The total cost savings range from 9.65% to 67.61%. The improvement increases as *ρ* increases. |

• With regard to departments where *ρ <* 1, the improvement mainly originates from reducing the labor expense. This result confirms the assumption that releasing all the capacity in these departments is thoroughly unnecessary. As for the department where *ρ* ≥ 1, cutting the waiting cost contributes most to the improvement. The waiting time has remarkably declined due to the shorter booking window design.

VII. CONCLUSION

Faced with the imbalance between healthcare supply and demand, timely access to doctors has become a major problem. This paper develops new stochastic models for designing and managing outpatient interday appointment systems that consider patients cancellation behaviors. We combine the following three system design parameters as appointment rules: appointment booking window, slot capacity, and service time of one block. Our models are intended to evaluate and design optimal appointment rules that balance resource utilization, patient access probability, and indirect waiting time.

We develop a novel renewal process to model the interday appointment system that engages appointment and potential patients. Compared with models in previous studies, our model can capture the potential waiting time, which is a critical indicator of system congestion in a limited booking window system. We use techniques, including the embedded Markov chain, the generating function approach, the rootsbased method, and the Laplace transform, to obtain the analytical steady-state distribution of our renewal model. This part of theoretical demonstration enhances the integrity and practicality of our model. Our model can therefore be applied to other fields of service management.

Our results indicate that system traffic coefficient is the main factor that decides how appointment rules can be made. For a system with sufficient supply, a small booking window is appropriate. The effect of the number of blocks is negligible. For a system with insufficient supply, sparse segmentation of daily blocks can reduce the system congestion probability and exerts an insignificant effect on resource utilization and patients waiting time. Appropriately lengthening the booking window can reduce the congestion probability and increase utilization.

Future research can focus on the following aspects. First, in this paper, the cancellation behaviors of patients are modeled as a binomial distribution with a constant cancellation rate for tractability. In our future work, we may extend this idea and pay attention to patients’ physician and time preferences that will affect their cancellation behavior. Second, we suppose that the service time and capacity for each block are equal. In future work, we will consider an even more flexible appointment rule.

APPENDIX A

PROOF OF COROLLARY 1

*Proof:* According to the one-step transition probability matrix P = {*p(i,j)A*}, it could be easily find that each state is accessible from the others, and the reverse is true, and so the states of the embedded Markov chain only has one class, and all states communicate with each other. Thus, our model is satisfied the irreducible condition. Then, following the basic theorem of recurrent states in discrete Markov chain, we know that the states in an irreducible chain are either all transient or all recurrent. Define *δi* = *E*[*XA(tr*+1*)* − *XA(tr)*|*XA(tr)* = *i*] to be the expected increasing patients in one block when system initially has *i* patients. From Pakes’ Theorem (Pakes 1969), every state is ergodic if: 1) |*δi*| *<* ∞ for all *i* values and 2) lim*i*→∞ *δi <* 0. First, we have

⎧*λl* − *i, i* ∈ {0*,*1*,...,n*} *δi* = ⎪⎨*λl* − *n, i* ∈ {*n* + 1*,... N* − 1 } (31)

⎪⎩*N* − *n* + *λl* − *i, i* ∈ {*N, N* + 1*,...*}*.*

For condition 1), all the terms on the right-hand side of (31) are bounded, and hence first case in (31) is satisfied. To demonstrate condition 2), we note that *δi* is definitely less than zero if *i* → ∞. So, the embedded Markov chain is irreducible and aperiodic, and we can write the balance equation of it for *i* ≥ *N*

⎛ *n* ⎞ *N*

*qiA* = *αi* ⎝*qj*⎠ + *αi*−*(j*−*n)qj j*=0 *j*=*n*+1

⎛ ∞ ⎞

|  |  |
| --- | --- |
| +*αi*−*(N*−*n)* ⎝ *qj*⎠*, i* ∈ {*N, N* + 1*,...*}*.*  *j*=*N*  Since lim*i*→∞ *αi* = 0, hence  ⎛ *n* ⎞ *N*  lim→∞ *qiA* = *i*lim→∞ *αi* ⎝*j*=0 *qj*⎠ + *i*lim→∞ *j*=*n*+1 *αi*−*(j*−*n)qj i* | (32) |
| ⎛ ∞ ⎞  + *i*lim→∞ *αi*−*(N*−*n) j*=*N qj*⎠ = 0*.*  ⎝ | (33) |



APPENDIX B

PROOF OF COROLLARY 2

*Proof:* Since the probability generating function *PA(z)* =

∞*i*=0 *qpi(Ajz,ii)A*(*qzjA*complex*,i* with |,*z*|we≤multiply1) and botheachsides*qiA* of=

the∞*j*=second0 equation by *z* and get *qiA z i*  *p(j,i)AqjA*.

Summing each of these equations from *i* = 0 to ∞ gives

∞ *qiA zi* = ∞ *αiz i* ⎛*n qj*⎞⎠ + ∞ *N*−1 *αizi*+*j*−*n* · *qj*

⎝

*i*=0 *i*=0 *j*=0 *i*=0 *j*=*n*+1

+ ∞ *αiz i*+*N*− *n* ⎛⎝∞ *qj*⎞⎠ (34)

*i*=0 *j*=*N*

which is equivalent to

∞ ⎡ *n N*−1 ∞ ⎤ *PA(z)* = *αizi* ⎣*qj* + *z j*−*nqj* + *qjzN*−*n*⎦

*i*=0 *j*=0 *jj*=*N*

=

*n*

+

1

1



*N*

−

1

= *eλl(z*−1*)* ⎡⎣*n qj* + *n z jqj* + ∞ *qjz N*−*n*⎤⎦

*z*

*j*=0 *j*=*n*+1 *j*=*N*

1

= *λl(z*−1*)* ⎡⎣*n j* + *zn PA )* − *n* − ∞ *i*

*e q(z qizi*  *qiz j*=0 *i*=0 *i*=*N*

+ ∞ *qjz N*−*n*⎤⎦*. j*=*N*

After identical transformation, we can get the expression of the probability generating function

*PA(z)* = *in*=−01 *q iA (zi* =*N qiA (zN* − *zi)*

−

*z*

*n*

*)*

−

∞

*i*



1

−

*z*

*n*

*e*

*λ*

*l*

*(*

1

−

*z*

*)*

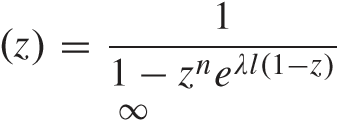
*.*



APPENDIX C

PROOF OF PROPOSITION 1

*Proof:* Equation (9) gives the expansion of denominator of our probability generating function in a power series. Let first obtain the expression of *biA* as follows:

*BA**,* |*z*| ≤ 1

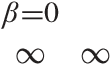


= *z nβ* · *eλlβ* · *e*−*λlzβ*



= *z nβ* · *eλlβ* · ∞ *(λlzβ)*! *γ*

*γ*

 *γ*=0

= *(λlβ)*! *γ e λlβznβ*+*γ.*

*γ*

*β*=0 *γ*=0

Thus

*BA(z)* = ∞ *ni*  *(*−*λ*−*lj)i*−*jn*! *e λljzi* (35)

*(i jn) i*=0 *j*=0

the coefficient of *zi* in series *BA(z)* equals to *biA* , which proves that *b*0*A* = 1 and

*biA* = *ni* *(*−*λ*−*lj)i*−*jn*! *e λlj.*

*(i jn) j*=0

Next, we rewrite the generating function as

*PA(z)* = ∞ *biA z i n*−1 *qiA (zi* − *zn)* − ∞ *qiA (zN* − *zi)*

*i*=0 *i*=0 *i*=*N*+1

hence

∞ *q A z i* = ∞ *biA zi n*−1 *qiA zi* − *n*−1 *qiA zn*

*i*

*i*=0 *i*=0 *i*=0 *i*=0

∞ *A N* + ∞ *qiA zi* *.* (36)

−

*qi z*

*i*=*N*+1 *i*=*N*+1

So, the coefficient of *zi* in the right-hand side after an expansion and removal of square brackets is just equal to *qiA*, but it needs a classified discussion according to the value of exponent in the power of right-hand side. At the beginning, we can obtain the recursion equation for *qiA*

*i*

|  |  |  |
| --- | --- | --- |
| *qiA* = *qjAb(i*−*j)*  *j*=0 | *A, i* ∈ {0*,...,n* − 1} | (37a) |
| =  *n*−1  *qiA* *qjAb(i*−*j) j*=0 | *n*−1  −  *b A(i*−*n)A qjA j*=0 |  |
| *n*−1  *qiA* = *qjAb(i*−*j)*  *j*=0  ∞  −*b* 0 *A*  *j* =*N*+1 | *i* ∈ {*n,..., N* − 1}  *n*−1  *A* − *b(i*−*n)A qjA*  *j*=0  *qjA, i* = *N* | (37b) (37c) |
| *n*−1 | *n*−1 |  |
| *qiA* = *qjAb(i*−*j)A* − *b(i*−*n)A qjA* | |

*j*=0 

∞

*b(i*−*N)A*  *qjA* + *qjAb(i*−*j)A j*=*N*+1 *j*=*N*+1

−

*i* ∈ {*N* + 1*,...*}*.* (37d)

Since *b*0*A* = 1 and *biA* = 0 for *i* ∈ {1*,...,n*−1}, the first class of *qiA* above is identical relation, which means the recursion formula of *qiA* for the first *n* could not be obtained. Therefore, there are three classes of *qiA* as proved earlier. 

APPENDIX D

PROOF OF COROLLARY 3

*Proof:* According to (11), we plug *b*0*A* = 1 into it and transpose the last term of right-hand side to left-hand side. It is easy to get the summation of all tail probabilities that

∞ *n*−1 *n*−1

*qiA* = *qjAb(N*−*j)A* − *b(N*−*n)A qjA. i* =*N j*=0 *j*=0

Because of (7), it provides us the normalization condition of *qiA* , and we can get the right-hand side of them by summing (37a), (37b), and (14)

∞ *N* min[*i,n*−1] *N n*−1

*qiA* = *qiAb(j*−*i)A* − *qiA b(j*−*n)A* (38) *i*=0 *j*=0 *i*=0 *j*=*n i*=0

and it equals to 1. Thus, we simplify the term of right-hand side by identity transformation

∞ *n*−1 ⎡ *n*−1−*i* ⎤ *n*−1 ⎡ *N*−*i* ⎤

*qiA* = ⎣*qiA*  *bβA*⎦ + ⎣*qiA*  *bβA*⎦ *i*=0 *i*=0 *β*=0 *i*=0 *β*=*n*−*i n*−1 ⎡ *N*−*n* ⎤ −⎣ *qiA*  *bβA*⎦ *i*=0 *β*=0

*n*−1 ⎡ *N*−*i* ⎤ *n*−1 ⎡ *N*−*n* ⎤

= ⎣ *qiA*  *bβA*⎦ − ⎣*qiA*  *bβA*⎦ *i*=0 *β*=0 *i*=0 *β*=0 *n*−1 ⎡ *N*−*i* ⎤

= *qiA*  *bβA*⎦

⎣ *i*=0 *β*=*N*−*n*+1

thus, we have

∞ *n*−1 *n*−1

*qiA* = *qiA b(N*−*j)A* = 1*.* (39) *i*=0 *i*=0 *j*=*i*



APPENDIX E

PROOF OF PROPOSITION 2

*Proof: R* is the coefficient determined by initial probabilities and *Q(z)* is a polynomial determined by tail probabilities whose exponents of *z* are greater or equal to *N*. Based on the inversion of the generating function, we have

*qiA* = *PA(i)*!*(*0*), i* ∈ {0*,*1*,...,n* − 1}*.* (40)

*i*

Define *HA(z)* = *nj**(z* −*z j)* = − *hiA zi* to be short, and the generating function could be written as

*PA(z)* = [*R* − *Q(z)*]*HA(z)BA(z).*

Considering *Q(z)*, it has common factor *zN*, so *Q(i)(*0*)* = 0, *i* ∈ {0*,...n*−1}. For *BA(z)*, with the expression of *biA* in (13), we find *BA(*0*)* = 1 and *B(Ai)(*0*)* = 0, *i* ∈ {1*,...n* − 1}. Thus,

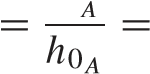
*PA(i)(z)* = *RHA(i)(z), i* ∈ {0*,*1*,...,n* − 1}*.* (41)

Since

*HA( i)(z)* = *ihtb*!*iA, i* ∈ {1*,...,n* − 1} (42) according to (40) and (42), we have

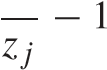
*qiA* = *RhiA , i* ∈ {0*,*1*,...,n* − 1}*.* (43)

Thus, the coefficient *γiA* in initial probabilities recursion formula *qiA* = *γiA q*0*A,i* ∈ {1*,...,n* − 1} equals to

*γiA* *h i*  *nh*−*iA*10 *z j*

*j*=

and

*n*−1 *γ A i* = *ni*=−*n*01 1*hiA zi* = *n* −1 *z .* (44) *i z*



−

*j*

=

0

*z*

*j*

*i*=0*j*=0



APPENDIX F

PROOF OF PROPOSITION 3

*Proof:* First, let *πi*∗*A (s)* and *di*∗*A (s)* defined as the Laplace–

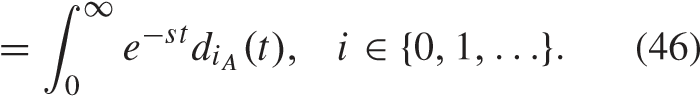
Stieltjes transform of *πiA (t)* and *diA(t)*, respectively, and

*L*

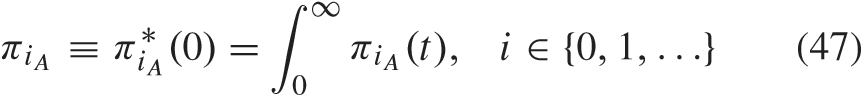
= ∞ *e* −*stπiA (t), i* ∈ {0*,*1*,...*} (45)

0

*L*



This follows from earlier that:





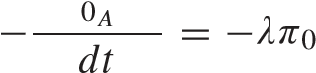
*diA* ≡ *diA (*0*)* = *diA (t), i* ∈ {0*,*1*,...*} *.* (48)

0

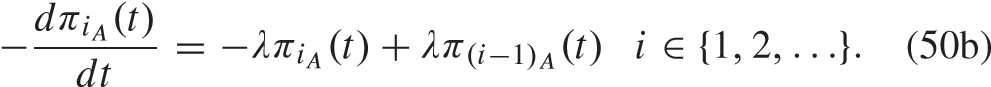
Also, based on the lag theorem of the Laplace transform, it could be obtained that

*L**.* (49)

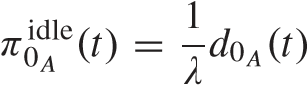
Note that for original renewal process, in the first period *t*[25], we have

*dπ (t)*

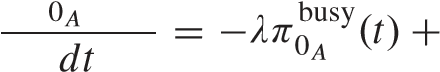
*A(t)* (50a)



For any other time within the period *t*

 (51)

*dπ*busy*(t) n*

−*djA(t* − *l)* (52)

*j*=1 *dπi (t)* busy

*diA(t)*

+*d(i*+*n)A(t* − *l), i* = 1 (53) *dπi (t)*

 *diA(t)*

+*d(i*+*n)A(t* − *l), i* ∈ {2*,..., N* − *n* − 1}

(54) *dπ* −*iA (t)* = −*λπiA (t)* + *λπ(i*−1*)A(t)* − *diA(t)*

*dt*

∞

+ *djA(t* − *l), i* = *N* − *n* (55)

*j*=*N*

*dπiA (t)*

*A(t)*

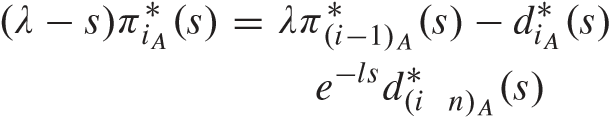
−*diA (t), i* ∈ {*N* − *n* + 1*,...*}*.* (56)

See [24] for details.

|  |  |
| --- | --- |
| Thus, the Laplace–Stieltjes transform of our differential equations (52)–(56) are  *n* | Using (59) to replace ∞*i*=1 *d*the*iA* (63)in (63)right-handleft-handside,sideweandget using (4) to identity transform  busy |

 *e*−*ls d*∗*jA(s)* (57a) 1*λ*−*lππ*0*A*0idle = *j*=*n* 1 *qjA. j*=1 *A*  *di*∗*A (s)* Thus

*), i* = 1 (57b) busy =  *n j*+=1 *qjA .* (64)

 *π*0*A λl q*0*A*

+ Since *π*0*A* = *π*0idle*A* + *π*0busy*A*

+

*i* ∈ {2*,..., N* − *n* − 1} (57c) *n*

 *j*=0 *qjA π* 0*A* = + *.* (65)

*λl q*0*A*

|  |  |  |
| --- | --- | --- |
| +*e*−*ls*  *d*∗*jA(s), j*=*N* | *i* = *N* − *n* (57d) | Based on (62b)–(62e), the expression of *πiA* , *i* ∈ {1*,*2*,...*} can be obtained recursively as follows: |

−

= − ∈ {− − + } *πi A* = *ij*=++ *qjA , i* ∈ {1*,*2*,..., N* − *n*} *i N n* 1*,... .* (57e)



∞

*j*

*i*

1



*λl q*0 *A*

Now adding (57a)–(57e), we obtain = =++ *qjA , i* ∈ {*N* − *n* + 1*,...*}*.* (66)

busy  ∞ *πi A λl q*0*A s π*0*A* ∗*(s)e A (s).* (58)

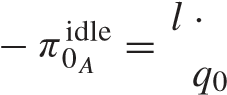


|  |  |
| --- | --- |
| *i*=1 *i*=1 |  |
| Taking limit as *s* → 0 and applying L’Hospital’s rule in lim*s*→0*(*1 − *e*−*ls/s)* = *l* yield  ∞  *π*0busy*A* +*iA* = *l diA.*  *i*=1 *i*=1  ∞  Since *πiA* = 1, so  *i*=0  ∞ | REFERENCES   1. L. V. Green and S. Savin, “Reducing delays for medical appointments: A queueing approach,” *Oper. Res.*, vol. 56, no. 6, pp. 1526–1538, 2008. 2. T. Cayirli and E. Veral, “Outpatient scheduling in health care: A review of literature,” *Prod. Oper. Manage.*, vol. 12, no. 4, pp. 519–549, 2003. 3. J. Feldman, N. Liu, H. Topaloglu, and S. Ziya, “Appointment scheduling under patient preference and no-show behavior,” *Oper. Res.*, vol. 62 , no. 4, pp. 794–811, 2014. 4. D. Gupta and B. Denton, “Appointment scheduling in health care: Challenges and opportunities,” *IIE Trans.*, vol. 40, no. 9, pp. 800–819 , |

 *l diA .* (59) 2008.

|  |  |
| --- | --- |
| *i*=1  According to the relationship between the departure distribution and the departure rate in (4) of Corollary 1, we yield | appointment-based services under patient no-shows,” *Prod. Oper. Manage.*, vol. 23, no. 12, pp. 2209–2223, 2014.  [6] N. Liu, “Optimal choice for appointment scheduling window under |

[5] N. Liu and S. Ziya, “Panel size and overbooking decisions for

 patient no-show behavior,” *Prod. Oper. Manage.*, vol. 25, no. 1 ,

*d* pp. 128–142, 2016.

1*.* (60)

[7] A. Ahmadi-Javid, Z. Jalali, and K. J. Klassen, “Outpatient appoint-

|  |  |
| --- | --- |
| Using (51) and eliminating *d*0*A*, we derive *q*  idle 0*A* | ment systems in healthcare: A review of optimization studies,” *Eur. J. Oper. Res.*, vol. 258, no. 1, pp. 3–34, 2017. [Online]. Available:  http://www.sciencedirect.com/science/article/pii/S0377221716305239  [8] C. Zacharias and M. Armony, “Joint panel sizing and appoint- |

*A*

|  |  |  |
| --- | --- | --- |
| 0 = +  *A λl q*0*A*  Next, setting *s* = 0 in (57a)–(57e), we get  *n*  *λπ*0busy*A* = −*d*0*A* + *djA*  *j*=0 | (62a) | ment scheduling in outpatient care,” *Manage. Sci.*, vol. 63, no. 11 , pp. 3978–3997, 2016.   1. J. Patrick, M. L. Puterman, and M. Queyranne, “Dynamic multipriority patient scheduling for a diagnostic resource,” *Oper. Res.*, vol. 56, no. 6 , pp. 1507–1525, 2008. 2. N. Liu, S. Ziya, and V. G. Kulkarni, “Dynamic scheduling of outpatient appointments under patient no-shows and cancellations,” *Manuf. Service* |

*π* *.* (61)

 *, i* 1 (62b) *Oper. Manage.*, vol. 12, no. 2, pp. 347–364, 2010.

|  |  |
| --- | --- |
| *λπiA* = *λπ(i*−1*)A* − *diA* + *d(i*+*n)A, i* ∈ {2*,..., N* − *n* − 1}  (62c)  ∞  *λπiA* = *λπ(i*−1*)A* − *diA* + *djA, i* = *N* − *n* (62d)  *j*=*N*  *λπiA* = *λπ(i*−1*)A* − *diA, i* ∈ {*N* − *n* + 1*,...*}*.* (62e) | ments with potential call-in patients,” *Prod. Oper. Manage.*, vol. 23 , no. 9, pp. 1522–1538, 2014.   1. L. W. Robinson and R. R. Chen, “A comparison of traditional and open-access policies for appointment scheduling,” *Manuf. Service Oper. Manage.*, vol. 12, no. 2, pp. 330–346, 2010. 2. J. Patrick, “A Markov decision model for determining optimal outpatient scheduling,” *Health Care Manage. Sci.*, vol. 15, no. 2, pp. 91–102, 2012. [14] N. T. J. Bailey, “On queueing processes with bulk service,” *J. Roy.* |

*A* = [11] R. R. Chen and L. W. Robinson, “Sequencing and scheduling appoint-

 *Statist. Society. Ser. B (Methodol.)*, vol. 16, no. 1, pp. 80–87, 1954.

Dividing (62a) by *diA* at both sides [15] N. K. Jaiswal, “Time-dependent solution of the bulk-service queuing problem,” *Oper. Res.*, vol. 8, no. 6, pp. 773–781, 1960.

*λπ*0busy *nj*=1 *djA .* (63) [16] R.*Sci.*Natarajan,*J.*, vol. “Discrete-time12, no. 4, pp.bulk317–326service, queuing1962. [processes,”Online]. Available: *Defence*

*A*



∞

*d*

*i*

=



∞

*i*=1 *A i*=1 *diA* http://publications.drdo.gov.in/ojs/index.php/dsj/article/view/6996

1. M. F. Neuts, “A general class of bulk queues with poisson input,” *Ann. Math. Statist.*, vol. 38, no. 3, pp. 759–770, 1967.
2. W. B. Powell, “Stochastic delays in transportation terminals: New results in the theory and application of bulk queues,” Ph.D. dissertation, Dept. Civil Eng., Massachusetts Inst. Technol., Cambridge, MA, USA, 1981. [Online]. Available: http://hdl.handle.net/1721.1/15840
3. M. L. Chaudhry, B. Madill, and G. Briere, “Computational analysis of steady-state probabilities of *M/G*[*a,b*]*/*1 and related nonbulk queues,” *Queueing Syst.*, vol. 2, no. 2, pp. 93–114, 1987.
4. M. L. Chaudhry, “Numerical issues in computing steady-state queueingtime distributions of single-server bulk-service queues: *M/Gb/*1 and *M/Gd/*1,” *ORSA J. Comput.*, vol. 4, no. 3, pp. 300–310, 1992.
5. M. L. Chaudhry and U. C. Gupta, “Modelling and analysis of

*M/G*[*a,b*]*/*1*/N* queue—A simple alternative approach,” *Queueing Syst.*, vol. 31, nos. 1–2, pp. 95–100, 1999.

1. O. Brun and J.-M. Garcia, “Analytical solution of finite capacity M/D/1 queues,” *J. Appl. Probab.*, vol. 37, no. 4, pp. 1092–1098, 2000.
2. G. Swartzman, “The patient arrival process in hospitals: Statistical analysis,” *Health Services Res.*, vol. 5, no. 4, p. 320, 1970.
3. Y. Bai, J. Song, and G. Zhang, “Design appointment in outpatient department with bulk service,” in *Proc. IEEE Int. Conf. Autom. Sci. Eng. (CASE)*, Aug. 2015, pp. 746–751.
4. J. Garcia, O. Brun, and D. Gauchard, “Transient analytical solution of M/D/1/*N* queues,” *J. Appl. Prob.*, vol. 39, no. 4, pp. 853–864, 2002.

**Jie Song** (SM’17) received the B.S. degree in applied mathematics from Peking University, Beijing, China, in 2004, and the M.S. and Ph.D. degrees in industrial engineering from Tsinghua University, Beijing, in 2007 and 2010, respectively. She is currently an Assistant Professor with the Department of Industrial Engineering and Management, Peking University. Her research interests include simulation optimization, and stochastic modeling in the application areas of logistics, healthcare, and production. She is a member of INFORMS.

**Yaqing Bai** received the Bachelor of Management degree from the Business School of Sichuan University, Chengdu, China, in 2012, and the Master of Management degree from the Department of Industrial Engineering and Management, Peking University, Beijing, China, in 2015. She is currently a Senior Business Intelligence Analyst with Amazon (CN), Beijing, China. Her research interests include stochastic modeling and queuing theory in service system, and particular in healthcare management.

**Jianpei Wen** received the B.S. degree in logistic engineering from the University of Science and Technology Beijing, Beijing, China, in 2013. He is currently pursuing the Ph.D. degree with the Department of Industrial Engineering and Management, Peking University, Beijing.

His research interests include healthcare operations, health policy, service operations, and rational queuing.

1. https://www.merritthawkins.com/2014-survey/patientwaittime.aspx 2https://en.m.wikipedia.org/wiki/Peking\_University\_Third\_Hospital [↑](#footnote-ref-1)