Expert Systems with Applications 41 (2014) 4529–4541

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Appointment scheduling algorithm considering routine and urgent patients

Jiafu Tang [a,b,*], Chongjun Yan [b], Pingping Cao [b]

[a] College of Management Science & Engineering, Dongbei University of Finance and Economic (DUFE), Dalian 116025, PR China

[b] Dept. of Systems Engineering, State Key Lab of Synthetical Automation of Process Industry, Northeastern University (NEU), Shenyang 110004, PR China

ARTICLE INFO

Keywords:
Appointment scheduling system
Health care operation management
Heuristics
Multimodularity
Two types of patients

ABSTRACT

This paper derives a solution approach to solve the outpatient appointment schedule problem for given numbers of routine and urgent patients considering a no-show probability to minimize the weighted sum of average patient wait time, physician idle time and overtime. An exact deterministic service time method is proposed to find the optimal schedule. An exponentially distributed service time property is presented to show that the objective function for routine and urgent patients is not multimodular, and consequently a local search algorithm based on multimodulary does not guarantee global optimality. Thus, a heuristic algorithm based on two kinds of shifting policies (HE-TKS) is developed to solve the appointment schedule, which gives a local optimal solution as an upper bound for the optimal schedule. Numerical experiments are conducted to illustrate how the critical factors affect service efficiency of the clinic in practice. It reveals that lower no-show probability, smaller interval lengths, shorter service times, and more urgent patients will benefit both patients and clinics.

1. Introduction

An appointment scheduling system needs to develop effective policies and decision tools to optimize clinics performance criteria, such as the average wait time, idle time, and overtime (Denton & Gupta, 2003; Robinson & Chen, 2003; Camilo & Robert, 2012). Appointment scheduling systems have been studied for more than half a century since Bailey (1952) and have received much attention recently (Cardoen, Demeulemeester, & Beliën, 2009; Ceschia & Schaerf, 2011; Denton & Gupta, 2003; Pegden & Rosenshine, 1990; Qu, Rardin, Williams, & Willis, 2007; Qu & Shi, 2011). Two problems exist for a well-designed appointment scheduling system: no-show patients and balancing routine patients with urgent care patients. The problem of no-show patients is a significant problem that affects the physician's ability to be useful and increases wait time for patients. No-shows not only reduce the clinic's revenue, but also jeopardize timeliness of service, which in turn increases patient dissatisfaction, because the physician stay idle in some intervals while there is congestion in other intervals. Lacy et al. (2004) reports that the probability of a no-show is anywhere from 28% to 45% in spite of reminder calls, emails, transport accommodations, and financial penalties. Field studies indicate that the no-show rate fluctuates widely depending on the medical specialty and patient demographic. In addition to the problem of no-show patients, this paper also considers the balance between long appointments and the need to receive same-day care. Routine patients and urgent patients are two groups of patients who need to be served using shared resources of clinics. A well-designed appointment system can balance the time and resources allocated to routine and urgent patients.

The contributions of this research are several. This paper extends the models proposed by Robinson and Chen (2010) and Kaandorp and Koole (2007) for homogeneous patients to two types of patients: routine patients and urgent patients. Urgent patients are those who need to receive service on the day they make appointments; routine patients are those who scheduled their appointments a few days or weeks in advance. Taking into account no-show patients, an appointment scheduling model with routine and urgent patients under deterministic and exponentially distributed service times respectively is proposed to minimize the weighted sum of the average patient wait time, physician idle time and overtime. Firstly, the traditional appointment scheduling policy with a given number of routine patients proposed by Robinson and Chen (2010) is extended to include both routine and urgent patients, and an optimal schedule policy is proposed for the case of deterministic service time. Secondly, under an exponentially distributed environment, theoretical results are proposed to show that the objective function for routine and urgent patients is not

* Corresponding author.
E-mail addresses: jftang@mail.neu.edu.cn, tangjiafu@dufe.edu.cn (J. Tang), yanchongjun0607@163.com (C. Yan).

http://dx.doi.org/10.1016/j.eswa.2014.01.014
0957-4174/

multimodular, and consequently the local search algorithm based on multimodulary proposed by Kaandorp and Koole (2007) could not guarantee global optimality. To address this problem, a heuristic algorithm based on two kinds of shifting policies (HE-TKS) is developed to solve the appointment schedule problem with routine and urgent patients under exponential service time, which gives a local optimal solution as an upper bound. Thirdly, numerical experiments are conducted to illustrate the impacts of critical factors on the service efficiency of the clinic in practice. The experiments reveal that both patients and clinics benefit when the no-show probability, interval length, service time decrease and the number of urgent patients seen increases.

The reminder of the paper is organized as follows. Section 2 presents a literature review of appointment scheduling and the multimodularity of the discrete function widely used in the paper. Section 3 develops a model for an appointment system with a deterministic service time and proposes an optimal scheduling policy. Sections 4 and 5 discuss the model for an exponentially distributed service time and present a heuristic procedure based on two kinds of shifting policies to find an upper bound for the optimal solution. Numerical results on the stability of the algorithm and the impact of the input parameters on system performances are presented in Section 6. Section 7 summarizes our findings and indicates the directions for future research.

2. Literature review

Scheduling patients to a physician has been widely studied by researchers e.g. Pegden and Rosenshine (1990), Qu et al. (2007), Ceschia and Schaerf (2011), and Petrovic, Morshed, and Petrovic (2011) in operations management. Cayirli and Veral (2003) provide a comprehensive literature review of appointment scheduling, clarifying the key factor in outpatient clinic and classifying the research methodology into three categories. Gupta and Denton (2008) point out new challenges and opportunities in appointment scheduling in healthcare management including primary care clinic, specialty hospital and surgery scheduling. Recent research integrates other complicated elements such as a comparison of different scheduling systems (Laganga & Lawrence, 2012; Robinson & Chen, 2010), heterogeneous patterns of arrivals and service time (Klassen & Rohleder, 1996; Zeng, Turkcan, Lin, & Lawley, 2010), patient classification (Qu & Shi, 2011; Saremi, Jula, ElMekkawy, & Wang, 2013), and the consideration of capacity policies (Dobson, Hasija, & Pinker, 2011; Gupta & Wang, 2008; Qu et al., 2007). A brief survey of articles on three relevant aspects is as follows:

The multimodular function is widely investigated as part of a queueing system. It is first introduced by Hajeck (1985) when dealing with the optimal admission control problem. Altman, Gaujal, and Hordijk (2000) extend the relation between convexity and multimodularity to some convex subsets; they obtain general optimization results for a sequence of multimodular functions and illustrate the feasibility of the theory in admission control of a D/D/1 queue with fixed batch intervals. Koole and Sluis (2003) design a local search algorithm converging to a global optimal solution based on multimodularity when investigating a shift scheduling problem for call centers with a concave service level objective. Kaandorp and Koole (2007) use the same method to decide the number of patients for a fixed length of each time slot to balance patient wait time and physicians idleness and overtime under the assumptions of a given number of homogenous (i.e. routine) patients with a no-show probability and exponential service time, but only routine patients are considered; no same-day appointments are accepted. Zeng et al. (2010) show their objective of profit maximization is multimodular for homogeneous patients in terms of a no-show probability, and identify the properties of an optimal schedule for heterogeneous patients, however the penalty cost includes only the number of patients not seen at the end of each interval. This paper considers both the interests of the patients and of the clinic.

A number of previous studies take patient classification into account in the design of appointment systems. Cox et al. (1985) propose a series of methods to sequence new and return patients, resulting in a shorter queue length and higher utilization of physicians. Klassen and Rohleder (1996) point out that scheduling patients with a small variance in service time in the beginning of the session performs better than other approaches considered. However, neither considers urgent care patients who arrive on the appointment day. Cayirli et al. (2008) use patients classification for both sequencing and interval adjustment to improve physician idle time, overtime, and patient wait time without any trade-offs in different clinic environments characterized by walk-ins, no-shows and the ratio of new patients to return patients. Due to the complicated situation, they identify some practical rules by simulation without optimization. Dobson et al. (2011) formulate a stochastic model to examine the effect of reserving intervals for urgent patients on the average number of urgent patients that are not handled during normal time and the average queue length of routine patients. They provide upper and lower bounds of capacity reservation without a concrete schedule. Cayirli et al. (2012) use simulation and nonlinear regression to derive an universal dome appointment policy with adjustment procedure that perform better than the traditional rules in terms of patient wait time, physician idle time and overtime. They provide a decision support tool to decide the interval length, but from a practical point of view, it is more reasonable to schedule patients at a predetermined interval (e.g. 15 min each interval) because patients will be inconvenienced by appointment times like 8:05:24. Begen and Queyranne (2011), Denton and Gupta (2003), and Erdogan and Denton (2013) also consider the appointment scheduling problem to search the optimal interval length under different assumptions. This paper decides the number of patients in a predetermined interval structure. Saremi et al. (2013) consider multiple resources (e.g. staff, operating rooms, surgeons, and recovery beds) to address the scheduling of outpatient surgeries for a multistage operating room serving multiple patient types, and develop simulation-based optimization methods to minimize patient wait time, patient completion time, and number of surgery cancelations.

No-shows receive comprehensive attentions from practitioners and researchers in OR/IE. Muthuraman and Lawley (2008) develop a stochastic overbooking model to mitigate no-shows and present a sequential myopic scheduling policy with an objective function that captures patient wait time, physician overtime, and patient revenue. Chakraborty, Muthuraman, and Lawley (2010) relax the assumption of an exponential service time to a general service time distribution and develop a similar myopic algorithm with an optimal stopping criterion. Both papers are dynamic decision problems in which the arrival sequence plays an important role on the optimal schedule. Green and Savin (2008) formulate an appointment system with no-shows as a single server queuing system in which the customers have a state dependent probability of not being served and may rejoin the queue, and derive stationary distributions of queue size for both deterministic and exponential distribution service time, but urgent patients are not included. Liu et al. (2010) develop a heuristic dynamic policies that considers the relation between no-shows and indirect waiting time from the booking day to the appointment day for a multi-period appointment scheduling problem. The simulation results show their approach outperforms other benchmark policies, particularly when the workload is high. Laganga and Lawrence (2012) construct a flexible appointment scheduling model to mitigate the disruptive effect of no-shows and develop a fast and effective scheduling procedure that