European Journal of Operational Research 255 (2016) 809–821

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Stochastics and Statistics

Appointment sequencing: Why the Smallest-Variance-First rule may not be optimal

Qingxia Kong[a,*], Chung-Yee Lee[b], Chung-Piaw Teo[c], Zhichao Zheng[d]

[a] School of Business, Universidad Adolfo Ibáñez, Chile
[b] Department of Industrial Engineering and Logistics Management, Hong Kong University of Science and Technology, Hong Kong
[c] Department of Decision Sciences, NUS Business School, National University of Singapore, Singapore
[d] Lee Kong Chian School of Business, Singapore Management University, Singapore

ARTICLE INFO

Article history:
Received 24 January 2015
Accepted 5 June 2016
Available online 11 June 2016

Keywords:
OR in health services
Appointment Sequencing
Smallest-Variance-First Rule
Stochastic Ordering

ABSTRACT

We study the design of a healthcare appointment system with a single physician and a group of patients whose service durations are stochastic. The challenge is to find the optimal arrival sequence for a group of mixed patients such that the expected total cost of patient waiting time and physician overtime is minimized. While numerous simulation studies report that sequencing patients by increasing order of variance of service duration (Smallest-Variance-First or SVF rule) performs extremely well in many environments, analytical results on optimal sequencing are known only for two patients. In this paper, we shed light on why it is so difficult to prove the optimality of the SVF rule in general. We first assume that the appointment intervals are fixed according to a given template and analytically investigate the optimality of the SVF rule. In particular, we show that the optimality of the SVF rule depends on two important factors: the number of patients in the system and the shape of service time distributions. The SVF rule is more likely to be optimal if the service time distributions are more positively skewed, but this advantage gradually disappears as the number of patients increases. These results partly explain why the optimality of the SVF rule can only be proved for a small number of patients, and why in practice, the SVF rule is usually observed to be superior, since most empirical distributions of the service durations are positively skewed, like log-normal distributions. The insights obtained from our analytical model apply to more general settings, including the cases where the service durations follow log-normal distributions and the appointment intervals are optimized.

1. Introduction

We study the design of a healthcare appointment system with a single physician and a set of patients whose service durations are random. The physician plans to see these patients in an appointment session within a fixed time interval – for example, from 8 a.m. to 12 p.m. Patients arrive punctually at their appointed times. Due to uncertainty in service duration, however, the physician may serve patients later than the appointed time (incurs patient waiting time) and/or the physician may not finish serving all the patients by the end of the appointment session (incurs physician overtime). A typical appointment design problem contains two sets of decisions: order of arrivals and time of arrivals. We refer to the first set as sequencing decisions and the second set as scheduling decisions. The objective is to minimize the expected total cost of patient waiting time and physician overtime.

We focus on the optimal sequencing problem, i.e., we decide the arrival sequence of the patients so that expected total cost to the system is minimized and scheduling decisions – the appointment interval assigned to each patient – will be fixed. We assume that patients are heterogeneous and can be classified according to the mean and variability of their service durations. As reported in the literature (cf. Cayirli, Veral, & Rosen, 2008; Klassen & Rohleder, 1996), it is common policy in many outpatient appointment systems to assign an equal appointment interval to each patient, which is also consistent with our experience in various local clinics. In practice, however, patients fall into different classes, and service duration can vary significantly for each class. Therefore, system performance largely depends on patients' arrival sequence.

In considering the optimal sequencing problem, it has been widely conjectured that the Smallest-Variance-First (SVF) rule is optimal (cf. Gupta, 2007; Wang, 1999; Weiss, 1990). For small cases with only two patients, the SVF rule has been proved to be optimal

* Corresponding author. Tel.: +56 223311323.
E-mail addresses: q.kong@uai.cl, qingxia.kong@gmail.com (Q. Kong), cylee@ust.hk (C.-Y. Lee), bizteocp@nus.edu.sg (C.-P. Teo), danielzheng@smu.edu.sg (Z. Zheng).

http://dx.doi.org/10.1016/j.ejor.2016.06.004
0377-2217/

under exponential service time distributions (cf. Wang, 1999) and convex ordering (cf. Gupta, 2007). However, it is only a conjecture that the SVF rule is optimal for more than two patients under general service time distributions. The problem has been investigated over decades and there is still an on-going research that tries to solve the problem. For larger systems, this conjecture is mainly supported by extensive simulation results that compare different sequencing policies (cf. Cayirli et al., 2008; Klassen & Rohleder, 1996). The intuition is that a patient with larger variance is more likely to overrun the stipulated session, and by putting him/her last, no later arrivals will be affected by the propagating effect of unpredictable wait times. To the best of our knowledge, however, no analytical results have been reported on the optimal sequence for more than two patients under stochastic ordering. Recent results reported by Mak, Rong, and Zhang (2015) reinforce the superiority of the SVF rule; they demonstrate, using a class of distributionally robust optimization models (in which only the marginal mean and standard deviation of each consultation duration are known), that the SVF rule is in fact optimal in the worst-case distribution under certain technical conditions. However, those conditions may not hold in all settings, and the correlations in the worst-case distribution are quite extreme. Finding the optimal sequence in general settings remains an important open problem.

By analyzing the problem under several assumptions, we shed light on why efforts to prove the optimality of the SVF rule in general settings have been futile. We first introduce a deterministic variant of the appointment sequencing problem (when service durations are known) and prove that the deterministic version of the problem is already $\mathcal{NP}$-hard. We then assume the appointment intervals are fixed according to a given template and the service time distributions are symmetric around the given appointment intervals, and analytically investigate the optimality of the SVF rule. In particular, we identify two important factors that affect the optimality of the SVF rule: the number of patients and the shape of service time distributions. We show that a large number of patients is needed for the SVF rule to be outperformed, which explains why the optimality of the SVF rule can only be proved for a small number of patients. Furthermore, the SVF rule is more likely to be optimal if the service time distributions are positively skewed, which explains why in practice, the SVF rule is usually observed to be superior, since most empirical distributions of the service durations are positively skewed, like log-normal distributions (cf. May, Spangle, Strum, & Vargas, 2011). Based on these insights obtained from our analytical study, we construct some counterexamples that the SVF rule is not optimal. Furthermore, our numerical analysis shows that these insights hold even when the service time distributions are not symmetric and the appointment intervals are optimized rather than fixed.

In this paper, we classify patients into distinct groups according to the characteristics of their consultation durations. We focus mainly on the impact of sequencing rules on patient waiting time and physician overtime. Our main contributions in this paper are as follows:

- We study a deterministic variant of the appointment sequencing problem, and show that even if each patient's consultation duration is deterministic (but may not coincide with the allocated appointment interval for that patient), the optimal deterministic sequencing problem is already $\mathcal{NP}$-hard. This result partly explains why finding an optimal sequence for the stochastic problem is so difficult.
- We use the theory of stochastic ordering to study the appointment sequencing problem. We obtain insights as to why scheduling patients using the SVF rule may not be optimal, and on the other hand, why in practice, the SVF rule is usually observed to be superior.
- We exploit the insights obtained from our analytical model to construct counterexamples showing that the SVF rule is not optimal. We show that these insights hold even when our model assumptions are relaxed, including the case where the appointment intervals are variable and optimized.
- Using likelihood ratio ordering, we show that the SVF rule is optimal for the last two patients. We also obtain several sufficient conditions under which the SVF rule is optimal.

2. Literature review

In appointment system design, studies tend to examine either scheduling or sequencing rules or the combination of the two. For the sake of brevity, we will not discuss the literature in detail here; interested readers are referred to Cayirli and Veral (2003), Gupta (2007), and Gupta and Denton (2008) for excellent reviews of healthcare appointment scheduling and sequencing problem.

Some studies assume that patients are homogeneous and use the First-Call-First-Appointment (FCFA) rule. Under these assumptions, scheduling rules (i.e., determining appointment intervals) are the main concern. There is extensive research being dedicated to optimal scheduling problem for a given sequence, and we briefly review a few here. Wang (1993) studied static and dynamic scheduling rule using queuing theory. Denton and Gupta (2003) developed a sequential bounding approach to determine the upper bounds of the problem. Begen and Queyranne (2011) showed that the scheduling problem can be solved in polynomial time when the cost function is linear and service durations follow discrete distributions. Ge, Wan, Wang, and Zhang (2014) extended their work to piecewise linear cost functions that are more practical. Cayirli, Yang, and Quek (2012) introduced a general scheduling rule that incorporates patient no show and walks-in and fits all clinic environments. In our paper, we use the methodology developed in Kong, Lee, Teo, and Zheng (2013) to solve for the near-optimal schedules when constructing counterexamples. Detailed discussion the methodology is included in Section 6 when we present our counterexamples. It is worthwhile to note that most methodologies developed for the optimal scheduling problem are capable of handling heterogeneous patients, but limited structural insights on the optimal schedules are available if patients are not homogeneous.

In practice, patients are distinct based on ages, type of procedure, nature of the visit, etc., and we can use variability in service duration to generalize these differences. Higher variance in service duration and/or a larger percentage of new patients will create higher variability in the system, and sequencing patients will be more valuable (cf. Cayirli et al., 2008; Robinson & Chen, 2003; Vanden Bosch & Dietz, 2000). Starting with Bailey (1952), this problem has been extensively studied over the past 60 years. In what follows, we will discuss the most relevant research on appointment sequencing problem.

Weiss (1990) was arguably the first to study the optimal sequencing problem analytically. In his 1990 paper, he jointly explored the optimal starting time and sequencing of surgical procedures to best utilize medical resources such as surgeons and operating rooms. Weiss showed that sequencing lower-variance procedures first is optimal in the case of two procedures under exponential or uniform service duration. He also conjectured that the SVF rule is optimal in general. Later on, similar results were reported for location-scale distribution such as normal and uniform distribution (cf. Gupta, 2007). Wang (1999) investigated the optimal appointment sequencing of n customers under exponential service distribution and proved the optimality of the SVF rule as Weiss (1990) did for $n = 2$ but used a different method. Wang argued that the result can be generalized to the case of n patients ($n > 2$) using a similar approach without any proof.