European Journal of Operational Research 237 (2014) 1142–1154

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Innovative Applications of O.R.

Computationally efficient evaluation of appointment schedules in health care

Stijn De Vuyst[a,*], Herwig Bruneel[b], Dieter Fiems[b]

[a] Department of Industrial Management, Ghent University, Technologiepark 903, B-9052 Zwijnaarde, Gent, Belgium
[b] Department of Telecommunications and Information Processing, Ghent University, St.-Pietersnieuwstraat 41, B-9000 Gent, Belgium

ARTICLE INFO

Article history:
Received 21 October 2011
Accepted 27 February 2014
Available online 15 March 2014

Keywords:
Stochastic Programming
Scheduling
Queueing
Complexity theory

ABSTRACT

We consider the problem of evaluating and constructing appointment schedules for patients in a health care facility where a single physician treats patients in a consecutive manner, as is common for general practitioners, clinics and for outpatients in hospitals. Specifically, given a fixed-length session during which a physician sees K patients, each patient has to be given an appointment time during this session in advance. Optimising a schedule with respect to patient waiting times, physician idle times, session overtime, etc. usually requires a heuristic search method involving a huge number of repeated schedule evaluations. Hence, our aim is to obtain accurate predictions at very low computational cost. This is achieved by (1) using Lindley's recursion to allow for explicit expressions and (2) choosing a discrete-time (slotted) setting to make those expressions easy to compute. We assume general, possibly distinct, distributions for the patients' consultation times, which allows to account for multiple treatment types, emergencies and patient no-shows. The moments of waiting and idle times are obtained and the computational complexity of the algorithm is discussed. Additionally, we calculate the schedule's performance in between appointments in order to assist a sequential scheduling strategy.

1. Introduction

1.1. Situation and scope

Because of its social and economic interest, the problem of scheduling a hospital's outpatients into the consultation session of a physician has received a lot of attention over the last sixty years. Many studies are motivated from a specific practical situation and aim at improving the organisational procedures in a particular (part of a) hospital (Babes & Sarma, 1991; Reinus et al., 2000; Zonderland, Boer, Boucherie, de Roode, & van Kleef, 2009; Harper & Gamlin, 2003). Clearly, practical settings differ considerably in terms of medical practice, organisation, regulations, administrative demands or limitations, preferences of patients or medical staff, management issues, etc. However, very often the underlying problem is largely the same and can be formulated as follows. Consider the practice of a physician who consults patients during a time interval of a certain length called a session, for example a 4-hour session from 8 am to 12 am every weekday. The physician is assisted by a nurse or secretary at the administration desk who is responsible for taking the calls of patients who wish to see the physician during the session of a particular day. The administrator must decide whether a calling patient can be admitted to that session and if so, at what time during the session the patient should arrive, i.e. what the appointment time of the patient is. All appointments are fixed before the session starts. The physician arrives at some point during the session, which is not necessarily the beginning. Given the session's length and the number of patients, a 'schedule' consists of both the patients' appointment times and the physician's arrival time.

Since patients are consulted one by one in their appointed order, the patients in the waiting room behave as a FIFO (First-In First-Out) queuing system with the physician as service facility. The time required to serve a single patient is the consultation time, comprising all actions by the physician devoted only to that patient such as examination, looking up test results, giving advice, writing prescriptions, updating files, and discussions. Prior to the session, consultation times are known stochastically only but can be assumed to be independent. The arrival process consists of scheduled patient arrivals at deterministic time points. Hence, evaluating a session amounts to the study of a queuing system conditioned on a certain sample path for the arrivals, which is known as an appointment system (Hassin & Mendel, 2008). A patient arriving to the session at its appointed time can encounter two possible

* Corresponding author. Tel.: +32 92645506.
E-mail addresses: stijn.devuyst@ugent.be (S. De Vuyst), hb@telin.ugent.be (H. Bruneel), df@telin.ugent.be (D. Fiems).

http://dx.doi.org/10.1016/j.ejor.2014.02.058

situations: either the physician has finished the consultations of previous patients or he has not. In the former case the physician has been without work since the departure of the last patient, whereas in the latter case it is the new patient who has to wait. As such, for each appointment there is either an idle time for the physician or a waiting time for the patient. As long as there is uncertainty on the consultation times when making the schedule, it is impossible to avoid both idle and waiting times although they can be controlled to a large extent by changing the schedule. In short, the aim of this paper is twofold.

- First, under rather general conditions, we devise an algorithm for evaluating a given schedule based on the consultation time distributions of the patients. The algorithm is in discrete time and produces the mean and variance of the waiting and idle times as a performance prediction. Its complexity is kept minimal by computing only the *change* of these moments from one patient to the next.
- Second, we visualise the schedule's 'virtual' performance in order to assist sequential scheduling of patients. That is, at any time point we calculate (mean and variance of) the so-called remaining work and running idle time. These virtual values become 'real' if an additional patient were to be scheduled at that time.

1.2. Related work

Depending on the specific situation, there are several so-called *environmental* factors that can make modelling the appointment systems considerably more complex, see Cayirli and Veral (2003) for an elaborate discussion. Patients without appointment may show up during the session ('walk-ins') but have to be seen by the physician anyway, either immediately (emergencies), in between regular patients or at the end of the session. Conversely, some patients that have an appointment do not show up for their consultation ('no-shows') or cancel the appointment too late. The no-show probability in some cases is up to 30%, depending on the type of health care offered and the patient population (Green & Savin, 2008; Sola-Vera et al., 2008; Lehmann, Aebi, Lehmann, Olivet, & Stalder, 2007). Clearly, walk-ins and no-shows contribute significantly to respectively the waiting and idle times of the schedule and to its overall uncertainty. Additionally, patients are not always punctual, for example arriving to the session later or sooner than they are supposed to. According to Alexopoulos, Goldsman, Fontanesi, Kopald, and Wilson (2008) the difference between appointed and actual arrival time is best modelled by an asymmetric Johnson distribution. Depending on the particularities of the used waiting-room policy, unpunctuality can result in overtaking of patients so that the original order of consultations is no longer maintained. With regard to scheduling, a complicating factor is also the fact that many patients have particular constraints concerning their appointment time. It is reported that as much as 25% of the calling patients ask to be given an appointment in a certain subset of the session (Rohleder & Klassen, 2000). As to which distribution is suitable for modelling patient consultation times, several propositions have been made. Originally, in Bailey (1952, 1955), Gamma distributions were used, as also preferred in e.g. Chakraborty, Muthuraman, and Lawley (2010). Other proposed distributions are Cox-type (Wang, 1997; Griffiths, Williams, & Wood, 2013), lognormal (Cayirli, Veral, & Rosen, 2008), Weibull (Babes & Sarma, 1991), uniform and/or exponential (Hassin & Mendel, 2008; Ho & Lau, 1992; Liu & Liu, 1998; Kaandorp & Koole, 2007) and even deterministic consultations (Green & Savin, 2008). However, patients may also be considered heterogeneous, i.e. have different consultation time distributions. Unlike walk-ins and no-shows, heterogeneity can reduce schedule uncertainty if properly taken into account. For each calling patient, the administration can estimate the required consultation time distribution based on the person's characteristics (age, medical record) and required type of medical treatment (medical scans, surgical procedures, inoculations, revalidation therapy, in-takes, discussion of test results, etc.).

For a more general overview of OR methods in health care planning and appointment scheduling we can refer to e.g. Brailsford and Vissers (2011), Cardoen, Demeulemeester, and Beliën (2010), Ivatts and Millard (2002), and Gupta and Denton (2008). Specific studies however can be divided into three classes based on the evaluation methodology. The first two rely on analytic methods and results from queuing theory, respectively in steady-state or transient, while the third class uses simulation. First, many of the queuing-theoretic approaches are based on classical results for the behaviour of a queue in steady-state. This is particularly useful for studying patient dynamics in larger health care facilities over long periods of time (weeks, months) and if one is interested in overall waiting time statistics rather than of individual patients. Historically, queueing models are classified using Kendall's notation $A/B/c/K$, where A characterises the interarrival time distribution, B the service time distribution, c is the number of servers and K is the capacity of the queue, i.e. the size of the waiting room. If $K = \infty$, it is omitted. Both fields A and B can e.g. assume the values M (exponential distribution), D (fixed value), PH (phase-type distribution) or G (general unspecified distribution).

Bailey (Bailey, 1954; Bailey, 1956) is among the first to advocate the use of statistics and queuing theory for capacity planning in hospitals. Some more recent studies however are the following. The efficacy of the Erlang $M/M/c/c$ loss formulas to predict the patient rejection rate at an intensive care unit is shown in McManus, Long, Cooper, and Litvak (2004). A tandem queuing model of both the intensive care unit and the operating table is given in Dijk and Kortbeek, 2009 where bounds on the rejection rate are derived. Similarly, in Gorunescu, McClean, and Millard (2002) an $M/PH/c/c$ model is used to quantify the number of beds necessary to meet a certain demand and in Tucker, Barone, Cecere, Blabey, and Rha (1999) an $M/M/1$ model suffices to assess operating room staffing needs during night shifts. Poisson tail probabilities in Vasanawala and Desser (2005) predict the required number of reserved slots (on a weekly basis) for emergency radiology given that 95% of the requests are accommodated. A preanesthesia evaluation clinic is reorganised in Zonderland et al. (2009) by applying an approximate decomposition method to an open multi-class queuing network of patients using the $M/G/1$ Pollaczek-Khintchine formulas. An $M/D/1/K$ and $M/M/1/K$ model with state-dependent no-shows is developed in Green and Savin (2008) to study the performance of 'advanced access' (Murray & Berwick, 2003), a recent paradigm where patients are offered same-day appointments as much as possible. In Reinus et al. (2000) a non-preemptive priority model is used to assess the average waiting times of emergent and non-emergent patients for computed tomography scans. In Creemers and Lambrecht (2009) matrix-analytic methods are used to analyse waiting times for an assignment system where patients can call in only during arrival sessions and are given an appointment during a service session. Arrival and service sessions are fixed and periodic. Assuming PH-type interarrival distributions, an algorithm is developed based on two hierarchical Markov chains with different time scales. In Creemers, Beliën, and Lambrecht (2012) the same authors use a bulk service queueing model for allocating slots to different patient classes. Finally, in Asaduzzaman and Chaussalet (2014) a network of perinatal care is modelled as a queueing network of $G/G/c/0$ loss systems.

All of the above studies assume an infinitely long queueing process of patients in equilibrium where all transient effects have subsided. Usually however, sessions are too short for steady-state predictions to be sufficiently accurate. Analysing a finite session