Article:

Ruiz-Hernández, D., García-Heredia, D., Delgado-Gómez, D. et al. (1 more author) (2019) A probabilistic patient scheduling model for reducing the number of no-shows. Journal of the Operational Research Society. ISSN 0160-5682

https://doi.org/10.1080/01605682.2019.1658552

A probabilistic patient scheduling model for reducing the number of no-shows

Diego Ruiz-Hernández[a] David García-Heredia[b] David Delgado-Gómez[b]
Enrique Baca-García[c]

[a]Sheffield University Management School, Sheffield, UK
[b]Department of Statistics, Universidad Carlos III de Madrid, Leganés, Spain
[c]Department of Psychiatry, Fundación Jiménez Díaz, Madrid, Spain

Abstract

Patients who do not attend their appointments, or "no-shows", cause the under-utilisation of the health centres' resources and increase the average waiting time for accessing specialty health care services. Although this problem has been addressed in different appointment scheduling models, behavioural issues associated to the patient's socio-demographic and economic characteristics and/or his or her diagnosis, have not been widely included in scheduling optimisation models. In this article, we propose an integer linear programming model thattakes into account such behavioural issues in order to reduce impact of no-shows in medical services. To achieve this goal, the objective function maximises the health centre's expected revenue by using show-up probabilities estimated for each combination of patient and appointment slot. These behaviour-based probabilities are obtained using both the individual's personal and clinical characteristics and his or her attendance history. In addition, the model takes into account the requirements imposed by both the health centre's management and the health authorities (e.g. distinguishing between first visits and follow-ups, among others). An extension of the model allows overbooking in some appointment slots. Experimental results show that the proposed model is capable of reducing the waiting list length by 13% and to attain an increase of about 5% in revenue when comparing to a basic model that assigns each patient to the first available slot. It was also observed that when overbooking was allowed in one to three slots per day, the waiting list was reduced between 30% and 62%; and the revenue increased by 7% to 13%.

Keywords: appointment scheduling; no-shows; overbooking; healthcare; behavioural OR

1 Introduction

Over the past decades, there has been a considerable increase on health care expenditures worldwide. For instance, in the United States, the percentage of the GDP spent on health has increased from the 12.51% in 2000 to the 16.84% in 2015 (World Bank, 2018). A not negligible part of this expense is caused by the patients, commonly referred to as no-shows, who do not show-up for their appointments. For example, Moore et al. (2001) concluded that "over the course of a year, total revenue shortfalls [due to no-shows] could range from 3% to 14% of total clinic income"; likewise, Berg et al. (2013) estimated daily losses of about 16.5% of the revenue for a no-show rate of about 18%. In overall, McKee (2014) estimates that no-shows cost the American healthcare industry around 150 billion dollars per year. No-shows have also an important negative effect on the efficiency of health systems, causing under-utilisation of resources, long waiting lists and decreased revenue. The volume of no-shows depends on elements as disparate as the region, the patient's socio-demographic characteristics, clinical diagnosis and prior no-show history, as well as the specialty and the type of service provided, among others (Dantas et al., 2018). In their literature review, Kheirkhah et al. (2015) refer reported no-show rates ranging from 3 to 80 percent. Along the same line, Moore et al. (2001) observed that no-shows and cancellations represent about 32.2% of scheduled time at a family planning residence clinic.

In order to reduce these figures, health centres utilise two alternative approaches. On one hand, the so-called active approaches include reminders and sanctions. The success of these methods is uncertain, with some research reports showing a drastic reduction in the percentage of no-shows after these measures are implemented (Molfenter, 2013), while others find no differences or, at most, a modest reduction (Hixon et al., 1999; Satiani et al., 2009).

This difference can be explained by the fact that the effectiveness of these methods may depend on the characteristics of the target population (Hashim et al., 2001). On the other hand, the so-called passive approaches aim at improving the current appointment system of the health centre by means of more sophisticated (and efficient) appointment assignment policies, instead of the most frequent practice of assigning the patient to the first available slot.

Optimising patient appointment systems has been an active subject of research over the last few decades (Cayirli and Veral, 2003; Gupta and Denton, 2008; Ahmadi-Javid et al., 2017). The patient allocation systems that have been proposed in the scientific literature present several differences, which are mainly consequence of the specific characteristics of the health centre and the type of service provided. For example, some centres establish that patients must receive their appointment at the time when this is requested, while in other cases appointments are scheduled at the end of certain period (the patient is notified later on by physical or electronic means). These two approaches are usually named as online and offline, respectively (Zacharias and Pinedo, 2014). Although online systems are the most frequently used, the rapid development of electronic appointment systems has caused an increase in the relevance of offline systems (Ahmadi-Javid et al., 2017). Another difference is whether the scheduling system admits overbookings or not, although most of the proposed systems include overbooking in their models (LaGanga and Lawrence, 2007; Chakraborty et al., 2010; Kim and Giachetti, 2006). A more detailed description of the different types of appointment systems can be found in the recent review conducted (Ahmadi-Javid et al., 2017).

Notwithstanding there is evidence that the probability that a patient will show-up to an appointment is closely related to his or her socio-demographic characteristics and condition (Dantas et al., 2018), traditional appointment scheduling models for medical services are usually based on the availability of slots, practitioner's timetables, and visit times, among other characteritics of the service provided. Only seldom, the proposed models take into consideration the probability that a patient will attend an appointment in a given time window. Moreover, those models tend to allocate probabilities based in generic data without taking into account characteristics and behavioural traits specific to each patient.

This article constitutes an effort for bringing the field of behavioural operational research to the area of patient schedulling, by proposing an appointment planning method that takes into account each individual's probabilities of no-show (estimated from their socio-demographic characteristics, diagnose and attendance history) for each specific combination of time-slot and patient. In doing so, our work seeks to fill a gap existing in the application of OR in healthcare (comprehensive reviews include those by Brailsford et al. (2009) and Hulshof et al. (2012)) through the development of behaviourally informed approaches (Hämäläinen et al., 2013), that aim at improving the provision of medical services by including associated patient's behaviour in the modelling process.

In this article, an integer linear programming (ILP) model is developed for optimising the offline assignment of medical appointments in a speciality service of a public health centre. The system aims at minimising the number of no-shows, and indirectly the waiting list length. This is attained by means of an objective function that maximising the expected revenue of the health centre. The model is designed as a single server system accounting for the fact that, in general, each practitioner has his own list of patients. Finally, as the health centre may be required by law to serve a fixed proportion of new patients every week, the model includes the possibility of reserving a percentage of slots for first-visits.

Under certain conditions, in order to reduce the large number of practitioners' idle periods caused by no-shows, a health centre may consider the possibility of introducing overbooking in some slots. This may also have a positive impact on the length of the waiting list (mainly in centres with large incidence of no-shows). For those cases, we propose a mixed integer linear programming (MILP) formulation that extends the initial model by allowing overbooking in a limited (pre-defined) number of slots.

Before introducing the mathematical formulation of the system, in Section 2, we provide a brief description of some related approaches available in the literature. In section 3 we present the proposed mathematical model. In Section 4 we conduct a simulation experiment in order to test our model's performance. We conclude this article in Section 5 with a discussion of the results and pointing out future lines of research.

2 Related literature

As mentioned above, several models have been proposed for improving patients' access to health care. The differences in these models are mainly consequence of the heterogeneity of the requirements imposed by the health centres (e.g. online or offline scheduling, single or multiple servers or if no-shows should be taken into account) and the goals pursued (e.g. maximise the revenue or reduce the length of the waiting list). In this section we focus our discussion on the analysis of those models most closely related to our work: first, we discuss the offline mathematical programming models (either ILP or MILP) proposed for single server systems; later, we present a review of some of the most relevant works that take into account the presence of no-shows from a probabilistic perspective.

Conforti, Guerreiro and Guido developed various ILP models that maximise the number of patients –weighted by the severity of their illnesses- scheduled for starting a radiotherapy treatment (Conforti et al., 2008, 2011). Their models assign each patient to several time slots during a given number of weeks so that the treatment can be conducted without interruptions. This assignment is conducted taking into account the constraints generated by patients that have already started the treatment. Zhu et al. developed a similar model for scheduling the access to a Magnetic Resonance Imaging scanner (Zhu et al., 2012). Their model assigns the patients to the required time slots in a two-week schedule so the number of allocated patients, weighted by their priorities, is maximised. Their model takes into account patients' time availability. Wang and Fung developed a model aiming at maximising profit, measured as the revenue earned from the attended patients minus the cost incurred from patients' rejection (Wang and Fung, 2014). The revenue was dependent on the patients' preferences for appointment time and practitioner. Additionally, a constraint was included for limiting the degree of discrepancy between the time allocated and the patient's preferences. More recently, Wiesche, Schacht and Werner proposed a MILP model that seeks to minimise the number of assigned appointments, penalising the number of patient shifted from morning to afternoon sessions (Wiesche et al., 2017). This allowed the authors, in one hand, to increase the time availability for attending walk-ins, and to balance the physicians' workload, on the other.

However, the models discussed above do not consider the existence of no-shows. In this regard, Savelbersbergh and Smilowitcz developed an ILP model whose objective function aimed at maximising the health condition of the population in a mobile asthma management program (Savelsbergh and Smilowitz, 2016). The health condition was measured by the likelihood that a patient's disease was controlled, which was strongly related to the probability that the patient showed-up to his appointment. The authors defined no-show probabilities for six different categories of patients depending on their preferences (strong or weak) for three different time windows (AM, noon, or PM) and 8 time slots in each time window. To our knowledge, this is the only offline ILP model that, although implicitly, takes into account the existence of no-shows.

Regarding the works that include no-show information from a probabilistic point of view, we find that most of them are developed from an on-line perspective and formulated as Stochastic Programming or Markov Decision Problems (Ahmadi-Javid et al., 2017). For example, Muthuraman and Lawley developed a stochastic overbooking model that considered each patient's no-show probability (Muthuraman and Lawley, 2008). The objective function aimed at maximising the revenue penalised by an overbooking cost, represented by the patient's waiting time and staff's overtime. This model was later tested by Daggy et al. on real data where the no-show probabilities were estimated applying a logistic regression to a dataset obtained from a Veterans Affairs medical centre (Daggy et al., 2010). In a different work, Glowacka, Henry and May estimated the probabilities that a patient will show-up to his or her appointment by means of an association rule mining technique (Glowacka et al., 2009). They used these probabilities to derive three manageable sets of rules for patient scheduling. Recently, Samorani and Laganga have proposed an online scheduling model that admits overbooking, and whose objective function aims at maximising the revenue penalised by the patients' waiting time and overtime cost. Instead of a probabilistic classifier, they use a binary one to maintain their problem computational tractable (Samorani and LaGanga, 2015).

The model proposed in this article extends the available literature in appointment scheduling for health centres in the following directions. First, unlike most of the mathematical programming-based research, our model takes into consideration the likelihood that a patient will not show-up to his or her appointment. Secondly, our formulation adopts an off-line approach that uses differentiated show-up profiles for each patient. These show-up profiles, that provide a specific show-up probability for each available slot, are obtained using socio-demographic and clinical characteristics of the patient. This is an important difference with respect to other available probabilistic work, which uses predominantly on-line approaches and/or where the no-show probabilities are either categorised (Savelsbergh and Smilowitz, 2016) or binarised (Samorani and LaGanga, 2015). A third characteristic is that, unlike other works that consider first visits and follow-ups as homogeneous groups or, plainly, ignore the first visit group (Daggy et al., 2010), our formulation distinguishes among them, allowing the model, apart from satisfying a legal requirement, to exploit the different characteristics of these groups. Finally, our model allocates priorities to the patients depending on the time they have been in the waiting list.

3 The Probabilistic Patient Scheduling Problem

In this section, we introduce a probabilistic scheduling model for reducing no-shows in specialty health centres that takes into consideration patient-specific probabilities of showing at each given day/time slot. The objective is maximising the centre's expected revenue by means of a reduction in the number of no-shows. The model distinguishes between two types of patients (first visits and follow-ups) and, by using a priority value associated to each patient, takes into account the time that the patient has remained in the waiting list. It also takes into account a Spanish legal constraint regarding the proportion of first visits that must be scheduled every week.

The following notation will be used in the mathematical formulation of the model.

Sets

$\mathcal{I}$, days of the week;

$\mathcal{J}$, time slots available in any given day;

$\mathcal{K}$, set of patients to be scheduled for appointment during the reference week.

Parameters

q, proportion of the number of available slots that must be allocated to first visits;

d_k, binary parameter indicating if patient $k \in \mathcal{K}$ has high ($d_k = 0$) or low ($d_k = 1$) priority during the reference week;

Z_k, binary parameter indicating if patient $k \in \mathcal{K}$ is a first visit ($Z_k = 1$) or a follow-up ($Z_k = 0$);

$P_{i,j,k}$, probability that patient $k \in \mathcal{K}$ will show-up to an appointment in slot $\{i, j\}$, for all $i \in \mathcal{I}$ and $j \in \mathcal{J}$;

w_z, revenue obtained either from a first visit ($z = 1$), or a follow-up ($z = 0$).

Variables

$X_{i,j,k}$, binary variable taking value 1 if patient $k \in \mathcal{K}$ is assigned to slot $\{i, j\}$, for all $i \in \mathcal{I}$ and $j \in \mathcal{J}$.

X_k^T, binary variable taking value 0 if patient $k \in \mathcal{K}$ is assigned a slot in the current week and 1 if the patient is referred back to the waiting list.

With this notation, and taking into account that the operator $\lceil \cdot \rceil$ rounds a real number to its upper integer value, the model is formulated as follows:

$$\max \quad \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} X_{i,j,k} P_{i,j,k} (Z_k w_1 + (1 - Z_k) w_0) \tag{1}$$

$$\text{s.t.} \quad \sum_{k \in \mathcal{K}} X_{i,j,k} \leq 1, \qquad \forall i \in \mathcal{I},\ j \in \mathcal{J} \tag{2}$$

$$\sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} X_{i,j,k} + X_k^T = 1, \qquad \forall k \in \mathcal{K} \tag{3}$$

$$\sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} \sum_{k \in \mathcal{K}} X_{i,j,k} Z_k \geq \min \Big\{ \sum_{k \in \mathcal{K}} Z_k, \lceil q |I| |J| \rceil \Big\}, \tag{4}$$

$$\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} X_{i,j,k}(1-d_k) \geq \min\Big\{\sum_{k\in\mathcal{K}}(1-d_k), \quad |I||J| - \min\big\{\sum_{k\in\mathcal{K}} Z_k, \lceil q|I||J|\rceil\big\}\Big\}, \tag{5}$$

$$X_{i,j,k},\ X_k^T \in \{0,1\}, \qquad \forall i\in\mathcal{I},\ j\in\mathcal{J},\ k\in\mathcal{K}. \tag{6}$$

The objective function maximises the cllinic's expected revenue. Notice that when $w_0 = w_1 = w$ the objective function maximises the expected showing-up rate; otherwise, the objective maximises the expected weighted showing-up rate. Constraints (2) guarantee that only one patient is assigned to each slot. Constraints (3) make sure that if a patient is not allocated in the current week, he or she is returned to the waiting list. As we are working with binary variables, constraints (3) also ensure that each patient is not allocated in more than one slot. Constraint (4) forces to reserve a number of slots for the first time visits. Constraint (5) guarantees that low priority patients will not be allocated to a slot as long as there are high priority patients unallocated.

Model with overbooking

As mentioned in the Introduction, there may be cases in which performing overbooking is considered convenient. For these situations, the baseline model is extended for allowing the possibility of assigning two patients to the same slot, provided that the sum of their showing-up probabilities is less than certain predetermined value. This is attained by introducing an overbooking penalty in the objective function and a number of associated constraints. The following additional notation is used in the extended model:

Parameters

C^{ov}, positive parameter representing the overbooking penalty;

M, constant satisfying $M > \max\{w_0, w_1\}$;

$G_{i,j}$, binary parameter taking value 1 if overbooking is allowed in slot $\{i,j\}$, for all $i\in\mathcal{I}$ and $j\in\mathcal{J}$;

$\pi_{i,j}$, parameter imposing a bound on the sum of the showing-up probabilities for any pair of patients simultaneously booked in slot $\{i, j\}$,for all $i \in \mathcal{I}$ and $j \in \mathcal{J}$ [1].

<u>*Variables*</u>

$Y_{i,j}$, binary variable taking value 1 if overbooking has been used in slot $\{i, j\}$, for all $i \in \mathcal{I}$ and $j \in \mathcal{J}$;

$O_{i,j}$, binary variable taking value 1 if at least one patient has been booked in slot $\{i, j\}$, for all $i \in \mathcal{I}$ and $j \in \mathcal{J}$.

The model with overbooking is then given by:

$$\max \quad \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}} \sum_{k\in\mathcal{K}} X_{i,j,k} P_{i,j,k} (Z_k w_1 + (1 - Z_k) w_0) - C^{ov} Y_{i,j} + M O_{i,j} \tag{7}$$

$$\text{s.t.} \quad \text{Constraints (3)-(5), and}$$

$$Y_{i,j} \leq G_{i,j}, \qquad \forall i \in \mathcal{I},\ j \in \mathcal{J} \tag{8}$$

$$\sum_{k\in\mathcal{K}} X_{i,j,k} \leq 1 + Y_{i,j}, \qquad \forall i \in \mathcal{I},\ j \in \mathcal{J} \tag{9}$$

$$\sum_{k\in\mathcal{K}} X_{i,j,k} P_{i,j,k} \leq \pi_{i,j}, \qquad \forall i \in \mathcal{I},\ j \in \mathcal{J} \tag{10}$$

$$O_{i,j} \leq \sum_{k\in\mathcal{K}} X_{i,j,k}, \qquad \forall i \in \mathcal{I},\ j \in \mathcal{J} \tag{11}$$

$$X_{i,j,k},\ X_k^T,\ Y_{i,j},\ O_{i,j} \in \{0, 1\}, \qquad \forall i \in \mathcal{I},\ j \in \mathcal{J},\ k \in \mathcal{K} \tag{12}$$

The $C^{ov} Y_{i,j}$ term in the extended objective function, equation (7), represents a penalty incurred when overbooking is used in slot $\{i, j\}$. Please notice that this term attains the largest possible reduction in the practitioners' idle times by allocating the same slot (if overbooking is admissible) to the pair of patients with highest sum of show-up probabilities. This is guaranteed by the fact that the larger the weighted sum of show-up probabilities, the larger the profit after discounting the overbooking cost for any given slot. Notice also that if $C^{ov} < \{w_0, w_1\} \times \min_{i,j,k} \{P_{i,j,k}\}$, the model will always use overbooking when $|\mathcal{K}| >$

[1]Notice that the probability of both patients showing-up is given by $P_{i,j,k} \cdot P_{i,j,k'}$, which attains a maximum at $P_{i,j,k} = P_{i,j,k'} = \frac{\pi_{ij}}{2}$ for any given value of π_{ij}.

$|\mathcal{I}||\mathcal{J}|$, i.e. whenever the number of patients in the waiting list is larger than the number of available slots. Likewise, if $C^{ov} > \max\{w_0, w_1\} \times \max_{i,j,k}\{P_{i,j,k}\}$, the model will never use overbooking.

Regarding the associated constraints, equations (8) define the slots where overbooking is allowed. Equations (9) limit the number of overbooked patients in a given slot to two, provided that overbooking is allowed. Finally, in order to control for the maximal probability of overcrowding (the case where two overbooked patients show-up for the same appointment), the sum of showing-up probabilities in an overbooked slot is bounded by parameter π_{ij} in constraints (10).

Term MO_{ij} in equation (7), together with constraints (11) and the fact that by definition $M > \max\{w_0, w_1\}$, ensures that the model does not consider overbooking unless all slots are used.

Additionally, our model presents the following two properties, which will be used in the computational implementation of the model for speeding up the execution:

Proposition 3.1. *In the model with overbooking, the $O_{i,j}$ variables always take integer values when they are relaxed to $0 \leq O_{i,j} \leq 1$ for all $i \in \mathcal{I},\ j \in \mathcal{J}$.*

Proof. Let O_{ij} be a continuous variable defined in the interval $[0,1]$ for all $i \in \mathcal{I},\ j \in \mathcal{J}$. If $\sum_{k\in\mathcal{K}} X_{i,j,k} = 0$, from constraint (11) it immediately follows that $O_{ij} = 0$. Alternatively, if $\sum_{k\in\mathcal{K}} X_{i,j,k} > 0$ and given that the X_{ijk} are binary variables, then O_{ij} can take any value in the interval $[0,1]$. However, given that MO_{ij} appears with positive sign in the objective function of the maximisation problem, it follows that $O_{ij} = 1$. □

Proposition 3.2. *In the model with overbooking, the $Y_{i,j}$ variables always take integer values when they are relaxed to $0 \leq Y_{i,j} \leq 1$ for all $i \in \mathcal{I},\ j \in \mathcal{J}$.*

Proof. Let $Y_{i,j}$ be a continuous variable defined in the interval $[0,1]$ for all $i \in \mathcal{I},\ j \in \mathcal{J}$. We consider two possible scenarios:

1. $|K| \leq |I||J|$: From the objective function if follows directly that $Y_{ij} = 0$ for all $i \in \mathcal{I},\ j \in \mathcal{J}$. Double booking any slot, when a number of slots remains unallocated, will imply unnecessarily incurring a penalty of C^{ov}.

2. $|K| > |I||J|$: Consider any given slot $\{i, j\}$. If overbooking is not allowed, $G_{i,j} = 0$, constraints (8) guarantee that $Y_{ij} = 0$.

Assume now that overbooking is allowed and conducted at some slot $\{i, j\}$, i.e. $G_{ij} = 1$ and $\sum_{k \in \mathcal{K}} X_{i,j,k} = 2$. Let $Y_{i,j} = \delta$ with $0 < \delta < 1$, satisfying constraints (8). From constraints (9) it follows that $\sum_{k \in \mathcal{K}} X_{i,j,k} \leq 1 + \delta$, and given that $X_{i,j,k} \in \{0, 1\}$ we conclude that $\sum_{k \in \mathcal{K}} X_{i,j,k} \leq 1$, which is a contradiction. Therefore, if slot $\{i, j\}$ is overbooked, then necessarily $Y_{ij} = 1$.

Finally, assume that overbooking is allowed but not conducted at some slot $\{i, j\}$, i.e. $G_{ij} = 1$ and $\sum_{k \in \mathcal{K}} X_{i,j,k} = 1$. Let $Y_{i,j} = \delta$ with $0 < \delta < 1$, satisfying constraints (8). As before, constraints (9) imply that $\sum_{k \in \mathcal{K}} X_{i,j,k} \leq 1 + \delta$, and given that $X_{i,j,k} \in \{0, 1\}$ it still holds that $\sum_{k \in \mathcal{K}} X_{i,j,k} \leq 1$. Now, given that $C^{ov} Y_{ij}$ appears with negative sign in the objective function of the maximisation problem, it follows that $Y_{ij} = 0$. Therefore, if slot $\{i, j\}$ is not overbooked, then immediately $Y_{ij} = 0$.

□

Comment

If instead of a penalty for overbooking, an expected cost for overcrowding was the driver behind the overbooking decision, the corresponding term in the objective function –and the associated constraints- will need to incorporate the overcrowding probability (the product of the attendance probabilities of the two overbooked patients). In this case, the objective function will seek to allocating the overbooked slots to the pair of patients with lowest product of show-up probabilities. Consequently, with the aim of minimising the overcrowding penalty, the system will still face large idle times (as the probability of none of the patients showing-up will still be large). Moreover, the problem will become non-linear.

3.1 Scheduling Procedure

The scheduling procedure works as follows:

1. A waiting list is available with the records of the patients waiting for appointment, including information about the number of weeks they have been in the list (sojourn)

and whether it is a first-time visit or not. New patients are added to the list at the time the appointment request is received and their sojourn length counter is initialised to zero.

2. The list of patients (henceforth referred to as the *buffer*) to be passed each week to the scheduler is built as follows:

   (a) The system first selects the patients with largest sojourn value and assigns them high priority ($d_k = 0$). This group contains both first-visits ($Z_k = 1$) and follow-ups ($Z_k = 0$).

   (b) Once the high priority patients have been selected, if the number of first-visits in the buffer is still below the legal requirement, the system sequentially adds first-visits in decreasing order of sojourn length until the requirement is satisfied or no more first visits are left in the waiting list. At each iteration, all first-visits in the corresponding sojourn level are included. These patients have low priority ($d_k = 1$) and $Z_k = 1$.

   (c) Finally, if after including high priority patients and first-visits, the number of patients in the buffer is smaller than the number of available slots (and there are still patients in the waiting list), the system sequentially adds patients in decreasing order of sojourn length until the size of the buffer is larger or equal to the number of available slots (or the waiting list is empty). At each iteration, all patients in the corresponding sojourn level are included. These patients have low priority ($d_k = 1$).

3. After this selection has been conducted, the system passes the list of candidates to the scheduler for solving the Probabilistic Patient Scheduling Problem with or without overbooking. Once the schedule has been obtained, the patients who did not receive an appointment are sent back to the waiting list with their original sojourn value.

Regarding the overbooking policy, whenever two patients show up for the same appointment, subsequent appointments are delayed until either a no-show happens and the last delayed patient takes that slot, in which case the original schedule is reestablished, or the day finishes and the practitioner does over-time until the list is cleared. Please notice that the over-time impact of this policy will be limited as long as the number of slots where overtime is admissible does not exceed a reasonable limit (e.g. no more than 2 or 3 slots).

4 Numerical Experiments

In order to evaluate the performance of our model, an experiment that reproduces the routine of a psychiatry department in a Spanish health centre was designed. In order to estimate the probabilities that the patients would show-up for their appointments, a database containing information from 47,118 visits to this department was used. In addition to the variable indicating whether the patient attended the appointment or not, this database contains several variables that have been frequently used to characterise non-shows. These variables were age (Alaeddini et al., 2011; Kopach et al., 2007), sex (Alaeddini et al., 2011), week day and time of the appointment (Glowacka et al., 2009; Daggy et al., 2010), lead time (time in queue) in weeks (Daggy et al., 2010), practitioner ID, appointment type (first visit or follow-up) (Kopach et al., 2007), number of previous appointments (Kopach et al., 2007; Daggy et al., 2010), and percentage of no-shows in previous appointments (Kopach et al., 2007; Daggy et al., 2010). The probabilities of show-up were obtained using a decision tree (Norris et al., 2014) classifier. The use of the database allowed us to obtain specific and differentiated attendance probabilities for each available appointment slot, provided the patient's profile. The simulation is conducted as follows:

1. At the beginning of each week, we generate the set of patients who call for a new appointment. To do this, a random number is generated according to a discrete uniform variable whose parameters are provided below. This number is used for randomly selecting a number of patients from our database. By doing this, we respect the proportion of first visits/follow-ups as well as the distribution of the variables representing the patients' characteristics. The selected patients are added-up at the end of the waiting list. Each patient in the waiting list has assigned a sojourn value representing the

number of weeks that he or she has remained in the list. New arrivals are all assigned a sojourn value equal to zero.

2. The list of patients to be passed each week to the scheduling model is built as described in item 2 of Section 3.1.

3. After this selection has been performed, the Probabilistic Patient Scheduling Problem is solved using the generated data[2]. Once the model makes the assignment, the parameters of the system are updated for the following week as follows:

   For each appointment we randomly determine whether the patient will show-up or not depending on the patient's estimated attendance probability given the allocated slot. If the patient shows-up to the appointment, the health centre obtains the corresponding income and the patient is removed from the system. Otherwise, the patient is either returned to the waiting list according to a predetermined probability, or eliminated from the system. If returned, the patient is put at the end of the list with sojourn value 0. This way, the experiment mimics the situation in which the patient that did not attend an appointment asks for a new one.

   Patients who did not receive an appointment are sent back to the waiting list with their initial sojourn time.

4. At the end of each scheduling stage, the sojourn values of all patients in the waiting list are increased by one.

4.1 Simulation framework

As we mentioned, our experiment reproduces the functioning of a psychiatry department week by week during one year (52 weeks). In this centre, the doctor does consultation from 8:30 to 15:30 from Monday to Friday and each consultation lasts 30 minutes. Therefore, if overbooking is not considered, the doctor would attend a maximum of 70 patients. Of those, at least 30% are first visits to comply with the regulatory requirements. For each first

[2] We solved the optimisation problems using Cplex 12.7.

visit the centre receives 70 euros and for each revision 50 euros. At the beginning of the simulation, it is assumed that there is a 7-week waiting list to access to the medical services. For each of the simulated weeks, the following operations are performed:

The weekly number of new requested appointments in the simulation follows a uniform $[51, 69]$ distribution. This choice, together with an estimated no-show rate of 24% and the 60% of no-shows who are referred back to the waiting list, returns an expected number of appointment requests of 68.64 per week. These figures guarantee that the weekly number of patients asking for a new appointment is always close to the 70 slots available for each practitioner.

Using this scenario, we test the following scheduling approaches:

1. The probabilistic scheduling model without overbooking.

2. The extended model with overbooking in three different situations: allowing overbooking just at 12:00 each day, allowing overbooking at 9:00 and at 12:00; and allowing overbooking at 9:00, at 10:00 and at 12:00. The reason why we chose these hours is because, in our database, they are the time-slots with the greatest number of no-shows. From hereafter, they will be referred as one, two and three rows of overbooking, respectively. In all of them, the parameter that limits the maximum expected number of patients $\pi_{i,j}$ is set to 1.5. Later, we will perform a sensitivity analysis to analyse the influence of this parameter.

3. The traditional model in which each patient is assigned to the first available slot (Daggy et al., 2010). We will refer this model as a FIFO system.

4.2 Results

Figure 1 shows the obtained results. In these plots, the legends “NoOver”, “Over1”, “Over2” and “Over3”stand for model without overbooking, and model with one, two and three rows of overbooking respectively.

Figure 1 (a) displays the number of people in the waiting list along the different weeks. It can be seen that the models which use overbooking obtain a fast reduction of the length

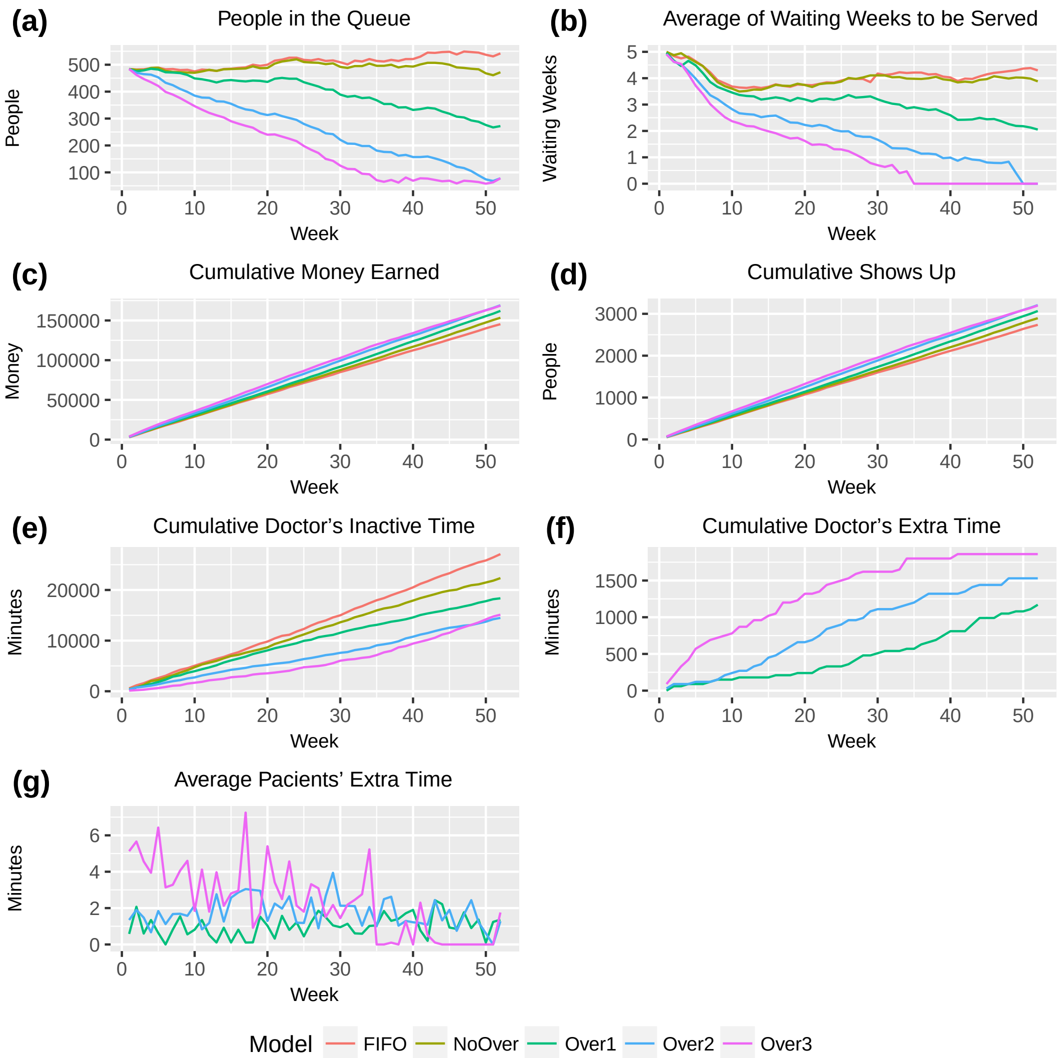

Figure 1: Simulation results.

of this list. It can be notice that the model which uses three rows of overbooking handles to eliminate the waiting list by week 35. After this week, the queue length is stable. One important result is that the proposed model that does not use overbooking (NoOver) is able to maintain the queue stable while the FIFO model cannot. At the end of the experiment, in week 52, the difference on the number of patients in the waiting list of these two models is 70 patients, which represents the complete schedule for a week. This result indicates that by only improving the patient assignment, without considering overbooking, is possible to avoid that the length of the waiting list increases.

Figure 1 (b) exhibits the mean time that the patients remain in the waiting list. These results are similar to the previous ones: the overbooking models reduce the mean time faster than the other two models, and the Over3 model stabilised around week 35. As before, the NoOver model attains to stabilised and the FIFO model do not. The drastic drop-out during the first weeks is consequence of the initial waiting list structure.

Figure 1 (c) shows the cumulative revenue. As expected, the models that have the greatest incomes are the overbooking models, followed by the NoOver and the FIFO models.

Figure 1 (d) illustrates the cumulative number of people who show up to the appointment. It can be noticed that this value is greater for the overbooking models. It is sensible to think that it is consequence of the fact that these models assign more patients, but it is also because these patients are optimally scheduled. This fact can also be appreciated in the NoOver and FIFO models. Despite they have the same number of assigned patients, the number of patients who show up is higher for the NoOVer model.

Figure 1 (e) shows the cumulative doctor's idle time. The first interesting fact is that this value is higher in the FIFO model that in the NoOver model even though doctors in these models have assigned the same number of patients. It was observed that FIFO had an average of 3.5 empty slots per day, while NoOver just 2.8. Regarding the overbooking models, it can be observed that, at the end of the simulation, models Over2 and Over3 have a similar cumulative doctor's idle time. This is, again, consequence that after week 35, the length of the waiting list for the Over3 model is minimal.

Figure 1 (f) displays doctor's overtime in which the effect of adding an extra row to the overbooking model can be appreciated. It is important to differentiate the curves before and after week 35 because, for the Over3 model, the waiting list is practically zero after this week as it was commented before. Therefore, after week 35, doctor barely suffer from over time in this model.

Figure 1 (g) represents the average time that each patient waited in the health centre to be attended. It can be observed that for the Over3 model, the patients have to wait between 2 and 6 minutes, which is a 6% and 20% of the time of each slot.

Next, a sensibility analysis is conducted in order to assess the effect of parameter π in the model's performance. To this end, the previous experiment is repeated for values $\{1.1, 1.3, 1.5, 1.7\}$ of this parameter. Table 1 shows the obtained results.

	Number in queue	Waiting weeks	Revenue (€)	Average show-ups (%)	Average empty slots	Average doctors' weekly overtime (min)	Average patients' extra waiting time (min)
FIFO	542	4.29	145,330	75.19	3.47	0	0
NoOver	472	3.88	153,430	79.53	2.86	0	0
Over1, $\pi_{i,j} = 1.1$	382	2.97	156,340	79.49	2.66	3.4	0.75
Over2, $\pi_{i,j} = 1.1$	262	1.90	161,200	77.69	2.29	5.1	2.35
Over3, $\pi_{i,j} = 1.1$	147	0.90	164,970	74.29	2.06	6.85	4.05
Over1, $\pi_{i,j} = 1.3$	280	2.14	161,990	78.48	2.33	18.85	3.95
Over2, $\pi_{i,j} = 1.3$	107	0.75	169,060	77.63	1.85	29.15	8
Over3, $\pi_{i,j} = 1.3$	84	0.15	168,190	76.40	1.91	33.45	9
Over1, $\pi_{i,j} = 1.5$	272	2.05	162,090	78.64	2.35	16.25	4.4
Over2, $\pi_{i,j} = 1.5$	78	0	169,140	77.50	1.85	36	9.6
Over3, $\pi_{i,j} = 1.5$	78	0	168,640	77.72	1.93	51.40	15.85
Over1, $\pi_{i,j} = 1.7$	263	1.87	163,270	79.20	2.26	20.55	5.05
Over2, $\pi_{i,j} = 1.7$	83	0.14	168,780	77.60	1.91	41.15	12.25
Over3, $\pi_{i,j} = 1.7$	83	0	168,530	77.40	2	63.5	20.05

Table 1: Results of the sensitivity analysis.

Some remarks on Table 1: i) values in the first three columns correspond to week 52; ii) values displayed in columns four and five are averages over 52 weeks; and iii) average values in the last two columns are calculated over the first 35 weeks to avoid the noise caused by the exhaustion of the waiting list (please see the comments around model Over3 earlier in this section). Moreover, for the sake of clarity, we report the average number of empty slots per day instead of doctor's idle time.

We notice that the model without overcrowding (NoOver) increases the centre's revenue in 5.5% with respect to the current policy (FIFO), reducing the waiting list by 13% in a year. The results show that a scheduling regime that assigns appointments taking into consideration the patient's characteristics may contribute –in the health centre under study- to a reduction of about 17.5% in the number of empty slots.

Regarding the overbooking model, the results depend on the value assigned to parameter π and the number of slots in which the overbooking is allowed ($G_{ij} = 1$). If, for instance, $\pi = 1.1$, it can be noticed that the impact on practitioners and patients is minimal. This is due to two main reasons: i) the small probability of overcrowding (two overbooked patients showing-up to the same appointment), 0.3 maximum; and ii) in the case of overcrowding, it occurs early enough for a no-show in later hours to compensate for the extra time devoted to attending the additional patient. For this value of π the revenue would increase in a range between 7% and 13% and the waiting list would be reduced from 30% to 72%. These values are consistent with the ones reported by Moore et al. (2001). We also notice that allowing overbooking always improves the health centre's revenue (with respect to the NoOver case), with the maximal revenue attained when overbooking is allowed in up to two slots (Over2). Moreover, allowing overbooking in two slots always reduces the number of empty slots, irrespectively of the value of π.

Finally, regarding the value of parameter π, the best results are obtained when this parameter takes values between 1.3 and 1.5. In those cases, the value of the objective function increases noticeably without imposing serious penalties on the patients, with average waiting times below 10 minutes for models Over1 and Over 2. These values return maximum overcrowding probabilities (two overbooked patients showing-up to the same appointment) of 0.42 and 0.56, respectively. This suggests that the optimal value of π should be such that the overcrowding probability is close to 0.5.

5 Conclusions

In this article we address the problem of no-shows in specialty clinics. This problem imposes large economic costs to the health centres –mainly due to practitioners' idle times-, and to the patients, who suffer the personal and economic impact of long waiting lists.

The no-shows problem is tackled in this article by proposing a scheduling strategy based on a mixed-integer programming model together with a dynamic priority allocation scheme. The proposed model aims at maximising the expected revenue of the health centre taking into account the revenue obtained from both first visit and follow-up patients. When the revenue of these two groups is the same, the objective function is equivalent to maximising the expected number of show-ups. The model takes into account several constraints imposed by both the law and the health centre's policies; among them, allocating a minimum percentage of the available slots to first visits, or assigning priorities based on the time the patient has been in the waiting list. Our formulation can be easily adapted for considering other types of priority, as jumping the queue when the severity of the patient's condition demands it,among others. The base model is extended for allowing the possibility of overbooking.

The maximisation of the expected number of show-ups is attained by using individualised show-up probabilities which depend on the patients' socio-demographic and personal characteristics as well as on his or her diagnosed pathology. These probabilities are computed for each day/slot combination using a decision tree classifier on a sample of nearly 50 thousand visits.

Simulation experiments show that whereas the waiting lists size increases on time when a FIFO scheduling regime is used, our base model is capable of reducing the waiting list and attaining a 5% increase in revenue with respect to the FIFO regime. Experimental results also suggest that a more significant reduction in the waiting list would be attained if overbooking was applied. The magnitude of this reduction would naturally depend on the amount of doctors' overtime that the health centre is willing to accept. It was observed that, by allowing overbooking in one time slot per day, a reduction of the waiting list of about

30% can be achieved at a minimum overtime cost. These results suggest a combined strategy where limited overbooking can be initially used for obtaining a significant reduction in the waiting list and, later on, switching to a regime without overbooking.

Our results point at two interesting research lines. The first one will aim at endogenising the number and selection of the appointment slots where overbooking is allowed. Given that not all the patients require the same consultation time, the second research line should extend the model for taking into account the expected consultation times of the different types of patient.

References

Ahmadi-Javid, A., Jalali, Z., and Klassen, K. J. (2017). Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research*, 258(1):3–34.

Alaeddini, A., Yang, K., Reddy, C., and Yu, S. (2011). A probabilistic model for predicting the probability of no-show in hospital appointments. *Health care management science*, 14(2):146–157.

Berg, B. P., Murr, M., Chermak, D., Woodall, J., Pignone, M., Sandler, R. S., and Denton, B. T. (2013). Estimating the cost of no-shows and evaluating the effects of mitigation strategies. *Medical Decision Making*, 33(8):976–985.

Brailsford, S. C., Harper, P. R., Patel, B., and Pitt, M. (2009). An analysis of the academic literature on simulation and modelling in health care. *Journal of simulation*, 3(3):130–140.

Cayirli, T. and Veral, E. (2003). Outpatient scheduling in health care: a review of literature. *Production and operations management*, 12(4):519–549.

Chakraborty, S., Muthuraman, K., and Lawley, M. (2010). Sequential clinical scheduling with patient no-shows and general service time distributions. *IIE Transactions*, 42(5):354–366.

Conforti, D., Guerriero, F., and Guido, R. (2008). Optimization models for radiotherapy patient scheduling. *4OR: A Quarterly Journal of Operations Research*, 6(3):263–278.

Conforti, D., Guerriero, F., Guido, R., and Veltri, M. (2011). An optimal decision-making approach for the management of radiotherapy patients. *OR Spectrum*, 33(1):123–148.

Daggy, J., Lawley, M., Willis, D., Thayer, D., Suelzer, C., DeLaurentis, P.-C., Turkcan, A., Chakraborty, S., and Sands, L. (2010). Using no-show modeling to improve clinic performance. *Health Informatics Journal*, 16(4):246–259.

Dantas, L. F., Fleck, J. L., Oliveira, F. L. C., and Hamacher, S. (2018). No-shows in appointment scheduling–a systematic literature review. *Health Policy*.

Glowacka, K. J., Henry, R. M., and May, J. H. (2009). A hybrid data mining/simulation approach for modelling outpatient no-shows in clinic scheduling. *Journal of the Operational Research Society*, 60(8):1056–1068.

Gupta, D. and Denton, B. (2008). Appointment scheduling in health care: Challenges and opportunities. *IIE transactions*, 40(9):800–819.

Hämäläinen, R. P., Luoma, J., and Saarinen, E. (2013). On the importance of behavioral operational research: The case of understanding and communicating about dynamic systems. *European Journal of Operational Research*, 228(3):623–634.

Hashim, M. J., Franks, P., and Fiscella, K. (2001). Effectiveness of telephone reminders in improving rate of appointments kept at an outpatient clinic: a randomized controlled trial. *The Journal of the American Board of Family Practice*, 14(3):193–196.

Hixon, A. L., Chapman, R. W., and Nuovo, J. (1999). Failure to keep clinic appointments: implications for residency education and productivity. *FAMILY MEDICINE-KANSAS CITY-*, 31:627–630.

Hulshof, P. J., Kortbeek, N., Boucherie, R. J., Hans, E. W., and Bakker, P. J. (2012). Taxonomic classification of planning decisions in health care: a structured review of the state of the art in or/ms. *Health systems*, 1(2):129–175.

Kheirkhah, P., Feng, Q., Travis, L. M., Tavakoli-Tabasi, S., and Sharafkhaneh, A. (2015). Prevalence, predictors and economic consequences of no-shows. *BMC health services research*, 16(1):13.

Kim, S. and Giachetti, R. E. (2006). A stochastic mathematical appointment overbooking model for healthcare providers to improve profits. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and humans*, 36(6):1211–1219.

Kopach, R., DeLaurentis, P.-C., Lawley, M., Muthuraman, K., Ozsen, L., Rardin, R., Wan, H., Intrevado, P., Qu, X., and Willis, D. (2007). Effects of clinical characteristics on successful open access scheduling. *Health care management science*, 10(2):111–124.

LaGanga, L. R. and Lawrence, S. R. (2007). Clinic overbooking to improve patient access and increase provider productivity. *Decision Sciences*, 38(2):251–276.

McKee, S. (2014). Measuring the cost of patient no-shows. *Power Your Practice*.

Molfenter, T. (2013). Reducing appointment no-shows: going from theory to practice. *Substance use & misuse*, 48(9):743–749.

Moore, C. G., Wilson-Witherspoon, P., and Probst, J. C. (2001). Time and money: effects of no-shows at a family practice residency clinic. *Family Medicine-Kansas City-*, 33(7):522–527.

Muthuraman, K. and Lawley, M. (2008). A stochastic overbooking model for outpatient clinical scheduling with no-shows. *IIE Transactions*, 40(9):820–837.

Norris, J. B., Kumar, C., Chand, S., Moskowitz, H., Shade, S. A., and Willis, D. R. (2014). An empirical investigation into factors affecting patient cancellations and no-shows at outpatient clinics. *Decision Support Systems*, 57:428–443.

Samorani, M. and LaGanga, L. R. (2015). Outpatient appointment scheduling given individual day-dependent no-show predictions. *European Journal of Operational Research*, 240(1):245–257.

Satiani, B., Miller, S., and Patel, D. (2009). No-show rates in the vascular laboratory: analysis and possible solutions. *Journal of Vascular and Interventional Radiology*, 20(1):87–91.

Savelsbergh, M. and Smilowitz, K. (2016). Stratified patient appointment scheduling for mobile community-based chronic disease management programs. *IIE Transactions on Healthcare Systems Engineering*, 6(2):65–78.

Wang, J. and Fung, Y. (2014). An integer programming formulation for outpatient scheduling with patient preference. *Industrial Engineering & Management Systems*, 13(2):193–202.

Wiesche, L., Schacht, M., and Werners, B. (2017). Strategies for interday appointment scheduling in primary care. *Health care management science*, 20(3):403–418.

World Bank (2018). Current health expenditure. https://data.worldbank.org.

Zacharias, C. and Pinedo, M. (2014). Appointment scheduling with no-shows and overbooking. *Production and Operations Management*, 23(5):788–801.

Zhu, H., Hou, M., Wang, C., and Zhou, M. (2012). An efficient outpatient scheduling approach. *IEEE Transactions on Automation science and engineering*, 9(4):701–709.