International Journal of Production Research

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tprs20

Optimising the booking horizon in healthcare clinics considering no-shows and cancellations

Gréanne Leeftink, Gabriela Martinez, Erwin W. Hans, Mustafa Y. Sir & Kalyan S. Pasupathy

To cite this article: Gréanne Leeftink, Gabriela Martinez, Erwin W. Hans, Mustafa Y. Sir & Kalyan S. Pasupathy (2021): Optimising the booking horizon in healthcare clinics considering no-shows and cancellations, International Journal of Production Research, DOI: 10.1080/00207543.2021.1913292

To link to this article: https://doi.org/10.1080/00207543.2021.1913292

Published online: 25 Apr 2021.

INTERNATIONAL JOURNAL OF PRODUCTION RESEARCH
https://doi.org/10.1080/00207543.2021.1913292

Taylor & Francis
Taylor & Francis Group

Optimising the booking horizon in healthcare clinics considering no-shows and cancellations

Gréanne Leeftink [a], Gabriela Martinez[b], Erwin W. Hans[a], Mustafa Y. Sir[b] and Kalyan S. Pasupathy[b]

[a]Center for Healthcare Operations Improvement and Research (CHOIR), University of Twente, the Netherlands; [b]Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, MN, USA

ABSTRACT

Patient no-shows and cancellations are a significant problem to healthcare clinics, as they compromise a clinic's efficiency. Therefore, it is important to account for both no-shows and cancellations into the design of appointment systems. To provide additional empirical evidence on no-show and cancellation behaviour, we assess outpatient clinic data from two healthcare providers in the USA and EU: no-show and cancellation rates increase with the scheduling interval, which is the number of days from the appointment creation to the date the appointment is scheduled for. We show the temporal cancellation behaviour for multiple scheduling intervals is bimodally distributed. To improve the efficiency of clinics at a tactical level of control, we determine the optimal booking horizon such that the impact of no-shows and cancellations through high scheduling intervals is minimised, against a cost of rejecting patients. Where the majority of the literature only includes a fixed no-show rate, we include both a cancellation rate and a time-dependent no-show rate. We propose an analytical queuing model with balking and reneging, to determine the optimal booking horizon. Simulation experiments show that the assumptions of this model are viable. Computational results demonstrate general applicability of our model by case studies of two hospitals.

ARTICLE HISTORY
Received 16 June 2020
Accepted 5 March 2021

KEYWORDS
Applications in healthcare systems; healthcare logistics; appointment scheduling; cancellation rate; queueing analysis.

1. Introduction

Healthcare services are continuously challenged to deliver efficient and effective patient care. Inefficiencies are among others caused by no-shows and cancellations. No-shows and cancellations not only result in adverse efficiency outcomes for clinics, but also in reduced quality of care for their patients (Davies et al. 2016). In order to mitigate the effects of no-shows and cancellations, these effects need to be incorporated in decisions on the design of appointment systems. This research therefore presents a data-driven queuing approach to account for no-show and cancellation behaviour in the design of optimal booking horizons for these clinics. The booking horizon determines how much time in advance an appointment can be planned, and is an input parameter to an appointment system. The challenge is that when the booking horizon is determined, there is no information on actual patient arrivals, as typically only historical data on the patient population is known. Therefore, the booking horizon optimisation problem is considered at the tactical level of control (Hans, Van Houdenhoven, and Hulshof 2012). Section 2 introduces what is known in the literature on no-shows and cancellations, which shows a need to include cancellations into non-attendance analyses and clinic design. In Section 3, we provide additional empirical evidence for incorporating no-show and cancellation behaviour in outpatient clinic design, by analysing the time-dependent behaviour of no-shows and cancellations based on real life data of two major healthcare institutions from the US and the Netherlands. Section 4 presents the queueing model that incorporates these no-shows and cancellations for determining the optimal booking horizon. Sections 5 and 6 present the numerical experiments and validation of the analytical model and results respectively. Section 7 gives the conclusions and discussion.

Our contribution is threefold: (1) We show the need and make the first step in incorporating time-dependent cancellations in outpatient clinic design. (2) Using data from two health systems in the US and the Netherlands we define the time-dependency of no-shows and cancellations, together with the timing of cancellations, and compare no-show and cancellation behaviour. (3) We develop and solve a data-driven queueing model

CONTACT A.G. Leeftink a.g.leeftink@utwente.nl

to determine the optimal booking horizon in which we are the first to take time-dependent no-shows and cancellations into account.

Throughout this manuscript we use the following definitions:

- *Cancellation interval*: The number of business days from the creation of an appointment to the date the appointment is cancelled.
- *Scheduling interval*: The number of business days from the creation of an appointment to the date the appointment is scheduled for.
- *Booking horizon*: The number of business days from the current date to the date of the latest available appointment slot.

2. Literature

2.1. Characteristics of no-shows and cancellations

Ever since the increasing focus on efficient healthcare operations, clinics started to evaluate their no-show and cancellation rates. No-shows and cancellations result amongst others in reduced productivity and efficiency for hospitals (Davies et al. 2016), financial impact through reduced revenue and idle resources (Moore, Wilson-Witherspoon, and Probst 2001; Norris et al. 2014; Bean and Talaga 1992), reduced learning opportunities for residents (Guse et al. 2003), and the waste of valuable resources, which could have been used to serve other patients. Furthermore, no-shows and cancellations increase the waiting lists, by reducing the number of appointments available. Therefore, it reduces patient access to care (Davies et al. 2016; Norris et al. 2014), which might affect the continuity of care for patients (Bean and Talaga 1992). Furthermore, the reduced patient access might cause a vicious cycle, with longer waiting lists increasing the non-attendance rates, which increases the waiting times again (Hawker 2007).

Although appointment attendance behaviour has been studied for over half a century, the high volume of recent medical research on this topic shows the problem is still present in healthcare systems. However, most of this literature only distinguishes between no-shows and shows, and excludes cancellations as a specific category from the analysis (Norris et al. 2014): Cancellations are either included as no-shows, included as shows, or excluded from the analysis all together. Only a few studies have analysed no-shows and cancellations as two separate conditions (Partin et al. 2016; Norris et al. 2014; Shah et al. 2016; Harris 2016), despite the different behaviour of patient cancellations compared to patient no-shows (Harris 2016). It is important to analyse patient cancellation behaviour as well, as cancelled appointments give opportunities to reallocate capacity (Norris et al. 2014; Harris 2016; Monahan and Fabbri 2018), and therefore to increase the clinic's utilisation and the number of patients that gets access to the clinic.

For clinics it might be challenging to fill appointment slots after last-minute cancellations, resulting in an idle resource, which has a similar effect as a no-show. Similar reasoning holds for patients that want to reschedule their appointment at late notice. To be able to assess this opportunity loss of cancelled patients, it is important to not only take the amount, but also the timing of cancellations into account. By quantifying this cancellation behaviour of patients, the effects of interventions can be measured, which is an open gap in the literature according to Monahan and Fabbri (2018). As an example, Chariatte et al. (2008) stated that in their healthcare institution there might be a peak in last-minute cancellations, by patients that want to avoid a payment for a missed appointment. We define the cancellation interval as the number of business days from the creation of an appointment to the date the appointment is cancelled. To the best of the authors' knowledge, data on the cancellation interval over multiple days is not reported before in the literature.

2.2. Scheduling characteristics as predictors of no-shows and cancellations

The relationship between the scheduling interval and the no-show and cancellation rates is well-studied. We define the scheduling interval, also referred to in the literature as lead time, planning horizon, appointment age, or appointment interval, as the number of business days from the creation of the appointment to the date the appointment is scheduled for. Focusing on predictive studies, Bean and Talaga (1992) and Norris et al. (2014) found that the scheduling interval is the most significant predictor of patient non-attendance, both for no-show and cancellation rates. Whittle et al. (2008) found a modest effect of the scheduling interval on no-shows, as for large scheduling interval the no-show rate stabilised. Furthermore, they found a highly significant effect of the scheduling interval on cancellations. Mohammadi et al. (2018) and Partin et al. (2016) found the scheduling interval to be a predictor of both no-shows and cancellations as well. Recently, machine learning techniques are employed to forecast no-show and cancellation behaviour. Denney, Coyne, and Rafiqi (2019) showed the scheduling interval was the top feature for predicting no-show and cancellations. Besides many studies that found a significant relation with the scheduling interval and

cancellations and/or no-shows (Davies et al. 2016), some studies did not find such a relation between the scheduling interval and no-show rate (Wang and Gupta 2011; Centorrino et al. 2001). Concluding, patients that have a longer scheduling interval tend to have a higher probability of no-show and cancellation. However, when the scheduling interval becomes very long, these effects may fade out (Bean and Talaga 1992; Whittle et al. 2008).

2.3. Strategies to minimising the effect of no-shows and cancellations

To mitigate the effects of no-shows and cancellations, clinics can try to influence patient behaviour or modify their scheduling strategy, for example through education, reminders, and financial rewards or penalties (Daggy et al. 2010).

Besides strategies to impact patient behaviour, scheduling strategies can be adopted. Scheduling strategies that aim to minimise the adverse effects of no-shows and cancellations include overbooking, open access scheduling, panel sizing, and reducing the booking horizon.

When *overbooking* is allowed, additional patients are booked to timeslots with a high probability of becoming idle, or booked in overtime, based on the probability that patients cancel or miss their appointment (Zacharias and Pinedo 2014). This way, the probability of resource idle time is minimised, and patients can get earlier access. However, overbooking may increase waiting times for patients that show up for their appointment, which could result in reduced patient satisfaction and lower attendance rates on the long term (Daggy et al. 2010).

Open access scheduling (also known as walk-in scheduling) schedules patients that require an appointment the same day, or allows patients to be seen at a walk-in basis (Robinson and Chen 2010). Since the scheduling interval is (close to) zero in this situation, the impact of cancellations and no-shows is small. However, high fluctuations in daily demand may result in idle and overtime. The hybrid policy, in which patients can both schedule an appointment or walk-in, allows using the idle time caused by no-shows to serve walk-in patients. However, Moore, Wilson-Witherspoon, and Probst (2001) showed that using a walk-in visit to cover idle time, does not lead to complete financial recovery, even if it leads to full utilisation.

Panel sizing limits the number of patients allowed in the patient panel, which includes all patients that can potentially use the service of the provider. This way, the waiting list can never explode, as the number of patients that can get admitted is controlled (Green and Savin 2008). Through the waiting list length, the number of no-shows and cancellations is controlled as well, as patients that are waiting longer have a higher no-show and cancellation probability. However, most outpatient clinics cannot limit their patient population, which makes this strategy especially valuable for the primary care setting.

By *limiting the booking horizon*, one can control the waiting list as well, and thus the number of no-shows and cancellations. However, rejecting all patients that require an appointment outside the booking horizon, might result in patient loss and under-utilisation of the system (Whittle et al. 2008). The booking horizon optimisation problem is a tactical level planning problem, which allows for taking on a higher-level methodology, such as queuing theory. Therefore, Liu (2016) developed an M/M/1/K queuing model, which penalises the patient loss, and considered a small revenue for empty slots, both due to under-utilisation and no-shows. This work is the closest to our proposed approach to optimise the booking horizon, but it only considers patient no-shows, and excludes patient cancellations.

Concluding, medical literature starts to recognise the need to include cancellations into non-attendance analyses, as cancelled appointments give opportunities to reallocate capacity (Norris et al. 2014). However, scheduling strategies and clinic designs do not take cancellations into account. As no-shows and cancellations depend on the scheduling interval, scheduling interval-dependent no-shows and cancellations should be taken into account in the design of outpatient clinics, whereas most literature assumes fixed cancellation and no-show rates (Ahmadi-Javid, Jalali, and Klassen 2016). In the remainder of this paper, we will first analyse the no-show and cancellation behaviour in varying outpatient clinics in two health systems, in order to define the scheduling interval-dependency of no-shows and cancellations, as well as the timing of cancellations. Furthermore, we advance the work of Liu (2016), by expanding their M/M/1/K queuing model with reneging in the queue, to incorporate both no-show and cancellation behaviour in outpatient clinic design.

3. Real life data analysis of no-shows and cancellations in US and the Netherlands

To design an appointment system that incorporates no-show and cancellation behaviour in Section 4, we need to get insight into this behaviour. Based on the literature analysis of Section 1, we hypothesise the no-shows and cancellation rates to depend on the scheduling interval. To show the practical need to include this time-dependent behaviour in the design of appointment systems in healthcare, this section presents applications from large medical centres in the US and the Netherlands.

The data collection is described in Section 3.1. Section 3.2 presents the no-show and cancellation outcomes. We summarise our results in Section 3.3.

3.1. Data sources

We included retrospective appointment scheduling data from two hospitals, namely Mayo Clinic in Rochester, MN, USA, and University Medical Center Utrecht in the Netherlands. These institutions will be referred to as Institution 1 and 2 in the remainder of the paper. Data of about 32,000 appointments was extracted from the hospital information system of Institution 1, and data of about 52,000 appointments was extracted from the hospital information system of Institution 2.

The data set of Institution 1 consists of almost 3 years of data (2014/01/01-2016/10/31), and includes all appointments that were scheduled during this time interval for one specialty. The data set of Institution 2 consists of 2 years of data (2015/01/01 – 2016/12/31), and includes all appointments that were scheduled in two outpatient clinics of two specialties. The outpatient clinics serve, among others, neurology, sexually transmitted diseases, and otorhinolaryngology patients, using an appointment system with fixed slot sizes. No walk-in patients are served in these outpatient clinics. Appointments are clustered in three categories, *Seen*, *Cancelled*, and *No-show*. Each appointment where a patient showed up is classified as Seen. When an appointment is cancelled or rescheduled more than 24 hours in advance, it is classified as Cancelled. Patients who are not present for their appointment without any notice, who are hospitalised, who are denied for service, and appointments that are cancelled or rescheduled within 24 hours of the actual appointment, are registered as a No-show. As both hospitals also have education and research tasks, we only included care related face-to-face appointments with a nurse practitioner or clinician.

3.2. No-show and cancellation rates

To analyse the institutions' no-show and cancellation rates, we perform several statistical tests, with the no-show and cancellation rates as dependent variables, and the scheduling and cancellation interval as independent variables. We calculate Spearman's ρ correlation coefficients to assess whether there is a monotonic relationship between appointment disposition and the scheduling interval. We perform a subgroup analysis for patient and clinic initiated cancellations, to evaluate whether the cancellation-motivation impacts our hospital data. To analyse the timing of cancellations, Spearman's ρ correlation coefficients are calculated to assess whether there is a monotonic decreasing relationship between appointment disposition and the scheduling interval. Furthermore, we perform a subgroup analysis for patients with various scheduling intervals to determine the timing of cancellations. We use IBM SPSS Statistics 22 for Windows for all statistical analyses.

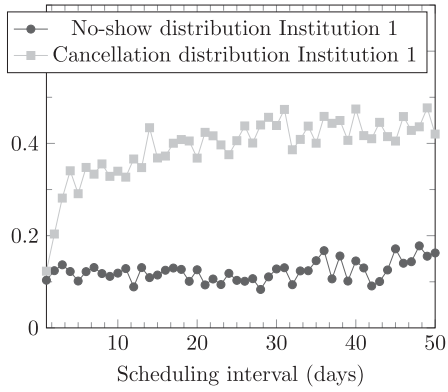


Figure 1. No-show and cancellation distributions per scheduling interval in days for the outpatient clinic of Institution 1.

3.2.1. Real life data based no-show and cancellation rates

For Institution 1, the no-show rate slightly increases from 10.3% for appointments that are scheduled the next day to 16.3% for appointments that are scheduled 50 days in advance (see Figure 1). A weak positive monotonic correlation is found between the daily lead time and the no-show rate (Spearman's $\rho = 0.344$, $n = 61$ working days, $p = 0.007$).

The cancellation rate increases from 12.3% for appointments that are scheduled the next day to 42.0% for appointments that are scheduled 50 days in advance (see Figure 1). A strong positive monotonic correlation is found between the daily lead time and the cancellation rate (Spearman's $\rho = 0.741$, $n = 61$ working days, $p < 0.001$).

For the first outpatient clinic of Institution 2, the no-show rate slightly increases from 9.1% for next day appointments to 11.0% for appointments that were scheduled 50 days in advance (see Figure 2). A weak positive monotonic correlation is found between the daily lead time and the no-show rate (Spearman's $\rho = 0.230$, $n = 61$ working days, $p = 0.075$).

The cancellation rate increases from 8.9% for next day appointments to 37.7% for appointments that were scheduled 50 days in advance (see Figure 2). A very strong positive monotonic correlation is found between the daily lead time and the cancellation rate (Spearman's $\rho = 0.877$, $n = 61$ working days, $p = 0.001$).

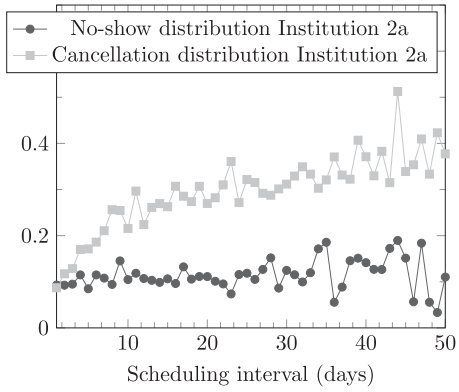


Figure 2. No-show and cancellation distributions per scheduling interval in days for the first outpatient clinic of Institution 2.

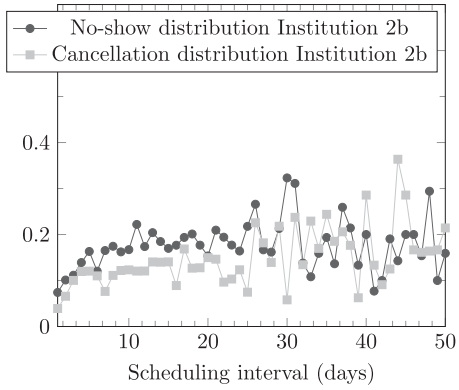


Figure 3. No-show and cancellation distributions per scheduling interval in days for the second outpatient clinic of Institution 2.

For the second outpatient clinic of Institution 2, the no-show rate slightly increases from 7.4% for next day appointments to 15.4% for appointments that were scheduled 50 days in advance (see Figure 3). A weak positive monotonic correlation is found between the daily lead time and the no-show rate (Spearman's $\rho = 0.301$, $n = 61$ working days, $p = 0.018$).

The cancellation rate increases from 3.9% for next day appointments to 21.4% for appointments that were scheduled 50 days in advance (see Figure 3). A moderate positive monotonic correlation is found between the daily lead time and the cancellation rate (Spearman's $\rho = 0.407$, $n = 61$ working days, $p = 0.001$).

3.2.2. Approximation of exponential distribution

In line with the literature (Green and Savin 2008; Liu 2016), Figures 1–3 show that the no-show and cancellation rates approximate an exponential distribution. Green and Savin (2008) propose the following no-show rate function:

$$\nu_j = \nu_{\max} - (\nu_{\max} - \nu_0)\exp^{\lfloor -j/\mu \rfloor / C},$$

Table 1. Parameter settings for no-show and cancellation rates per scheduling interval in days

	$\nu^a_{\max}$	ν^a_0	C	*Significance*
No-show rate Institution 1	0.137	0.083	13	$p \leq 0.001$
No-show rate Institution 2a	0.181	0.096	97	$p \leq 0.001$
No-show rate Institution 2b	0.189	0.000	3	$p \leq 0.001$
Cancellation rate Institution 1	0.457	0.000	5	$p \leq 0.001$
Cancellation rate Institution 2a	0.311	0.000	7	$p \leq 0.001$
Cancellation rate Institution 2b	0.137	0.000	3	$p \leq 0.001$

[a] For cancellation rates this reflects χ.

where $\nu_{\max}$ reflects the maximum observed no-show rate, ν_0 the minimum observed no-show rate, and C is a scaling parameter. As μ is the service rate and j the number of timeslots, j/μ is the number of days in the scheduling interval. Similar reasoning holds for the cancellation rate:

$$\chi_j = \chi_{\max} - (\chi_{\max} - \chi_0)\exp^{\lfloor -j/\mu \rfloor / C}.$$

We find the best-fit parameter values by minimising the sum of the mean squared errors between the observed and the modelled no-show and cancellation rates, to maximise the goodness of fit. This way we find a no-show rate and cancellation rate for each institution, which are displayed in Table 1, together with the statistical significance of the fitted distributions to the data (based on a χ-squared test).

3.2.3. Cancellation timing

The cancellation timing provides insight in the reuse potential of cancelled appointments slots (Monahan and Fabbri 2018). As no timing behaviour over multiple days is reported in the literature, we hypothesise that patients cancel their appointments both early and late in their scheduling interval, as they realise right after scheduling the appointment that a date is not convenient, or realise when the appointment is coming closer that for example other commitments are more important than this appointment. As we expect this behaviour to be more distinct for patients with larger scheduling intervals, Figures 4 and 5 show the cancellation timing behaviour for Institution 1 for various subgroups based on increased scheduling intervals (similar results for Institution 2 not shown). In Figure 5, we normalised the scheduling intervals on the interval [0, 1], with 0 being the date on which the appointment is created, and 1 the appointment date. As shown, the cancellation timing indeed follows a bimodal distribution, with a peak right after the create date of the appointment, and right before the appointment date. This indicates that independent of the scheduling interval, people tend to cancel their appointment either right after an appointment is made (about half of the cancelled appointments iscancelled within 5 working days), or right before the appointment

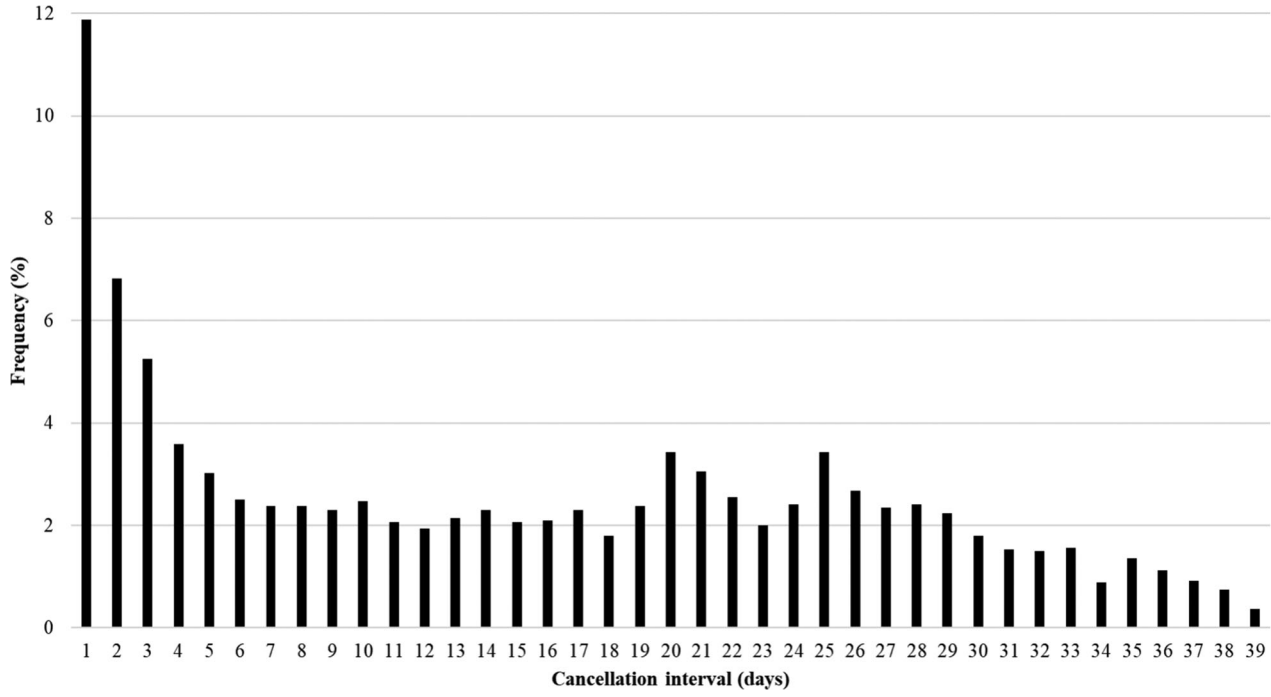


Figure 4. The fraction of cancelled appointments per cancellation interval for appointments with a scheduling interval of 20-40 days.

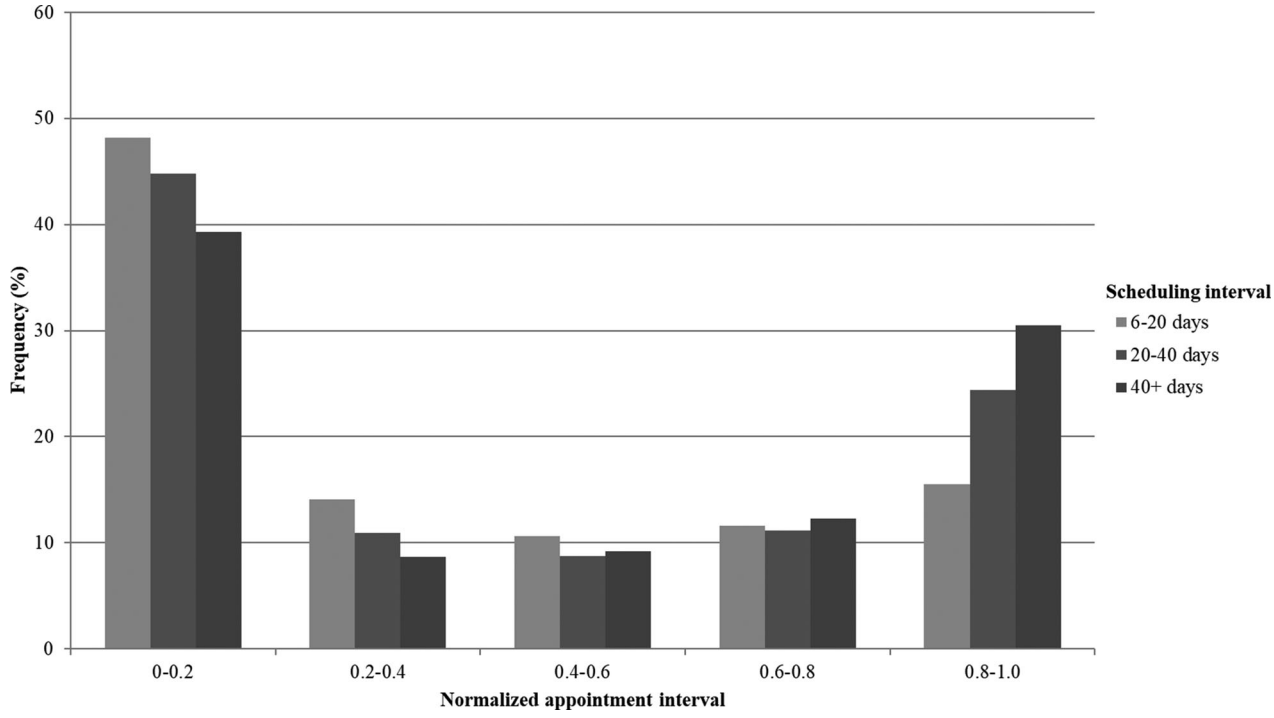


Figure 5. The probability of the timing of a cancellation for a given scheduling interval.

will otherwise take place (about two-third of the cancelled appointments is cancelled less than 5 working days before the actual appointment date). Thus, a patient that did not cancel in the first one-third of the scheduling interval, has a probability of cancellation of 8% in the middle part. An appointment that survived until the final third of the scheduling interval, has a cancellation probability of 14%, and a no-show probability on the day of the appointment of 12%. Note that the frequency plots for appointments scheduled within 5 days are not shown, as the bimodal behaviour is especially visible for cancellations with larger scheduling intervals.

For small scheduling intervals both peaks merge into one peak.

3.2.4. Initiation of cancellations

A cancellation occurs by patient or clinic initiation. As clinic initiated cancellations reflect system behaviour, these cancellations may behave differently. Therefore, Whittle et al. (2008) and Blæhr et al. (2016) performed two analyses for both patient initiated and clinic initiated cancellations. Both studies found significant relations and observed similar cancellation rate behaviour for patient and clinic initiated cancellation rates. Furthermore, Foreman and Hanna (2000) analysed the impact of the scheduling interval on attendance rates, and found that this impact is independent of the reasons for non-attendance. In line with this literature, we find that Institution 1 initiated 11% of the total cancellations, which shows the majority of cancellations is patient initiated. Institution 2 initiated 42% and 13% of its cancellations respectively. Reasons for the clinic to cancel the appointment are related to scheduling errors and unexpected changes in provider calendars due to for example illness. Summarising, both patient and clinic initiated cancellations show similar significant monotonic increasing behaviour and timing pattern.

3.3. Summary of the results

This section analysed the no-show and cancellation behaviour of two healthcare systems. We analysed both US and EU based outpatient clinics, and conclude that no-show and cancellation behaviour is similar for the various health systems, as monotonic increasing rates are observed, as well as bimodal cancellation timing behaviour.

This is the first study to analyse the timing of cancellations. We observe bimodal behaviour, with two cancellation peaks, right after the moment that the appointment is scheduled, and right before the actual appointment time. This is an important observation, as slots of appointments cancelled in the first peak can be reassigned with a high probability to new patients. However, slots of appointments cancelled in the second peak are less likely to be reassigned. This effect has to be taken into account in the design of appointment systems.

A comparison of the obtained no-show and cancellation rates shows that the no-show rate converges faster than the cancellation rate. This is in line with the literature (Whittle et al. 2008). Therefore, we expect that reducing the booking horizon has a larger influence on the cancellation behaviour of patients than on the no-show behaviour of patients.

Concluding, we observe scheduling interval dependent no-show and cancellation rates from US and EU practice. As this impacts the possible performance of an appointment system in clinics, these systems need to be designed and optimised taking the time-dependent behaviour into account.

4. Model

We focus on finding the optimal booking horizon, as this approach allows for implementation in outpatient clinical practice, and includes the time-dependency of no-shows and cancellations. The booking horizon problem is a tactical level planning problem on the organisation of healthcare delivery processes at an intermediate planning horizon (Hans, Van Houdenhoven, and Hulshof 2012). Therefore, we take on queuing theory, which is regarded as a higher-level methodology. The operational level planning, which focuses on day-to day processes such as appointment scheduling and will not likely reach a steady-state at any point during the day, is outside the scope of this research. For this same reason, our tactical analysis does not take into consideration operational behaviour, for example, variability in appointment duration, and wait-time patterns during the day.

The booking horizon can be expressed in the number of slots in the future in which appointments can be scheduled. The booking horizon is preferably set in such a way that the number of patients rejected because of unavailability of appointment slots, is minimised, while at the same time the number of patients served is maximised. To maximise the number of patients served, the effects of cancellation and no-show rates on idle slots and system capacity are minimised. As we are interested in numbers of patients served, patients rejected and idle slots, we can model this problem as a finite queue queuing system with reneging, which incorporates the monotonic behaviour of the cancellation and no-show rates analysed in Section 3. All future appointments in an appointment system can be together considered as the queue, which makes the booking horizon equal to the maximum queue capacity. By limiting the maximum queue capacity, we can evaluate the effects of a limited booking horizon on idle time, and the proportions of patients served and rejected. We validated the queuing model using the simulation-based approach described in Section 6. Our work differs from the previous queuing models analysed in Green and Savin (2008) and Liu (2016) as they do not consider cancellations, for they are excluded (Liu 2016) or included as no-shows (Green and Savin 2008). More specifically, we consider a single-server queuing system with no-shows, reneging in the queue, and balking to evaluate the optimal booking horizon. Patients are served

on a First Come First Serve (FCFS) basis, and due to the finite capacity of the appointment system, patients that arrive with $K-1$ patients in the system queue will leave. Cancellations are patients who leave the queue before their appointment. Furthermore, the system encounters no-shows. When a patient does not show-up for an appointment, the server will be empty for the entire service time of this patient. No overtime, and no preemption of service is allowed, and similar to Liu (2016) and Green and Savin (2008) we assume exponential service times.

In this study, we assume that patients are offered one appointment slot on a FCFS basis, and service times of appointments are exponentially distributed with mean time μ. Under these assumptions, the system can be modelled as a M/M/1/K queuing system with capacity K and μ appointment slots provided in a unit of time (Liu 2016).

Our goal is to find the booking horizon, which is equal to $\lfloor K/\mu \rfloor$, that maximises the appointment system revenue. This revenue is a combination of an added reward of serving patients and a penalty for rejecting patients.

We assume patients arrive from an infinite source according to a Poisson distribution with rate λ. Patients are served by a single server with exponential service rate μ. Patients do not enter the queue if they encounter a full queue at their arrival. Each patient is rejected at an opportunity cost θ_B. A patient always enters the queue if it is not full.

Each patient in the queue can cancel his/her appointment generating a cost θ_C, since this patient is lost. A patient waits a random amount of time before cancelling, which is assumed to have a negative-exponential distribution with constant rate α. The rate α represents the average number of cancellations of the system per unit of time (section 6 addresses the dependency of α on K). Consequently, the long-run probability that any one of the j patients scheduled in the system may cancel his/her appointment is equal to $c_{j+1} = j\alpha, j = 0, \ldots, K-1$ (Ancker and Gafarian 1962). Although the cancellation timing could be bimodal as shown in Section 3.2, the model focuses on cancellations that could occur close to the allocated slot. The simulation model in Section 6 further evaluates the impact of bimodal-cancellation timing.

Each patient that enters the queue and does not cancel before service, has a probability of not showing up for his/her appointment. The probability that a new arrival will be a no-show when upon arrival there are j patients scheduled in the queue is equal to ν_{j+1}. Based on the no-show rate behaviour analysed in Section 3, we can assume that the no-show probability of the system can be described by a monotonic sequence $\nu_{j-1} \leq \nu_j, j = 1, \ldots, K-1$. A patient that is served provides a nominal unit of revenue.

Let $p_j(K)$ be the steady-state probability that upon arrival there are j patients scheduled in the system, and $p_0(K)$ be the steady-state probability that the system is idle. Let $\rho = \frac{\lambda}{\alpha}$, $\delta = \frac{\mu}{\alpha}$, then the steady-state equations for the M/M/1/K queuing system are (Ancker and Gafarian 1962), for $j \in 0, \ldots, K-1$:

$$p_{j+1}(K) = \frac{\rho}{\delta + j} p_j(K), \quad \sum_{j=0}^{K} p_j(K) = 1.$$

Let $\Gamma(\cdot)$ be the gamma function defined as $\Gamma(z) = \int_0^{+\infty} t^{z-1} e^{-t}\, dt$, then the closed-form expressions of the steady-state probabilities are:

$$p_j(K) = \rho^j \frac{\Gamma(\delta)}{\Gamma(j+\delta)} p_0(K), \quad j = 1, \ldots, K, \tag{1}$$

$$p_0(K) = \frac{1}{1 + \Gamma(\delta) \sum_{j=1}^{K} \frac{\rho^j}{\Gamma(\delta+j)}}. \tag{2}$$

Let $P_S(K)$ be the proportion of patients served, $P_N(K)$ the proportion of no-show patients, $P_C(K)$ the proportion of cancellations, and $P_B(K)$ the proportion of blocked patients. We have the following expressions:

$$P_S(K) = \frac{\rho}{\delta}\left(1 - p_0(K)\right) - P_N(K),$$

$$P_C(K) = 1 - p_K(K) - \frac{\rho}{\delta}(1 - p_0(K)),$$

$$P_N(K) = \sum_{j=0}^{K-1} p_j(K)\beta_j\nu_j, \quad P_B(K) = p_K(K),$$

where $\beta_j = \frac{\delta}{\delta+j}$ is the probability that a new arrival that joins the queue does not cancel its appointment (Ancker and Gafarian 1962). Let $\theta_B \geq 0$, and $\theta_C \geq 0$, the long-run expected revenue of the system is defined as:

$$R(K) = \lambda P_S(K) - \lambda P_B(K)\theta_B - \lambda P_C(K)\theta_C. \tag{3}$$

The booking horizon problem can be formulated as follows:

$$\sup_{K \in \mathbb{Z}^+} R(K). \tag{4}$$

This problem can be easily solved by limiting the optimisation domain with a constant $\bar{K} \in \mathbb{Z}^+$, which depends on the queue parameters and the cost coefficients weights as shown in the Appendix.

5. Experiment settings and results

This section describes the numerical experiments. First, the base case and experiment settings are described in Section 5.1. Second, Section 5.2 presents the experimental results.

5.1. Base case and experiment settings

We consider an outpatient clinic which operates five days a week. Every day, six appointment slots are available. Weekends are excluded from the analysis. As six appointment slots are available per day, we set the deterministic service rate $\mu = 6$, and arrival rate $\lambda = 6$, with patients arriving according to a Poisson distribution. The no-show and cancellation rates are exponentially distributed, and derived from the data-analysis of Section 3. We consider the no-show rate of Gallucci, Swartz, and Hackerman (2005) (G05) as a no-show rate, as G05 has been used in the literature most frequently (Liu 2016; Green and Savin 2008). Furthermore, patients cancel their appointments with rate $\alpha = 0.06$, as derived from Institution 1 (see Figure 1).

As all cost coefficient weights are normalised towards the revenue from serving one patient in one timeslot, we need to assess the cost of cancellation (θ_C) and the cost of rejection (θ_B). We expect the cost of cancellation to be higher than the cost of rejection. As we expect rejected patients to be booked in another clinic, or be overbooked in non-clinic hours, which is the current practice in both hospitals included in this research, we do not consider a cost of lost patients for rejected patients, but we do include an inconvenience cost. Cancelled patients however might end up being lost by the clinic, as not every patient will reschedule their appointment. Furthermore, cancellations have a higher impact on the system (i.e. through an extra administrative burden, blocking slots for patients that would have showed up). As there is a tradeoff between rejection and cancellation, decision makers should together decide upon the cancellation and rejection cost coefficient weights, based on the aforementioned considerations. Therefore, we experiment with various cost coefficient weights as shown in Table 3. In the base case we use the settings $\theta_B = 1.2$ and $\theta_C = 1.4$.

To evaluate the efficiency of the method and to assess the behaviour of various system settings, we run the following experiments, as shown in Table 3. First, we analyse the impact of the no-show and cancellation rate on the optimal booking horizon. Eight different no-show rates are considered, five derived from the literature and three derived from hospital data (refer to Section 3). Although many studies report on the time dependency of no-show rates, most literature does not include a functional form of the time dependent no-show rate which is based on real-life data (Green and Savin 2008). Furthermore, most literature does not force their rates to long term asymptotic behaviour, despite the fact that both no-show and cancellation probabilities are not allowed to exceed one. Therefore, we limit our literature rate inclusion to rates that are monotonically increasing and converging towards a maximum value, which does not exceed one. We were able to identify five studies that provided such measurements over multiple scheduling intervals, from which we can derive a functional form, for which the parameters are presented in Table 2.

Table 2. Parameter settings for literature based no-show rates per scheduling interval in days

Study	Name	ν_{max}	ν_0	C
Benjamin-Bauman et al. 1984	BB84	0.48	0.16	7
Festinger et al. 2002	F02	0.67	0.05	2
Gallucci et al. 2005	G05	0.43	0.11	2
Green and Savin 2008	GS08	0.31	0.01	50
Whittle et al. 2008	W08	0.21	0.11	6

Table 3. Input parameter variations for the experiments

Exp no.	μ	λ	No-show rate	Canc. rate	(θ_B, θ_C)
Base case	6	6	G05	0.06	(1.2, 1.4)
1	6	6	BB84	0.06	(1.2, 1.4)
2	6	6	F02	0.06	(1.2, 1.4)
3	6	6	GS08	0.06	(1.2, 1.4)
4	6	6	W08	0.06	(1.2, 1.4)
5	6	6	Inst. 1	0.06	(1.2, 1.4)
6	6	6	Inst. 2a	0.06	(1.2, 1.4)
7	6	6	Inst. 2b	0.06	(1.2, 1.4)
8	6	6	G05	0.10	(1.2, 1.4)
9	6	6	G05	0.075	(1.2, 1.4)
10	6	6	G05	0.05	(1.2, 1.4)
11	6	6	G05	0.025	(1.2, 1.4)
12	6	5	G05	0.06	(1.2, 1.4)
13	6	7	G05	0.06	(1.2, 1.4)
14	6	8	G05	0.06	(1.2, 1.4)
15	6	10	G05	0.06	(1.2, 1.4)
16	6	6	G05	0.06	(1.1, 1.5)
17	6	6	G05	0.06	(0.8, 0.9)
18	6	6	G05	0.06	(0.8, 1.2)
19	6	6	G05	0.06	(1, 1)
20	6	6	Inst. 2a	0.093	(1.2, 1.4)
21	6	6	Inst. 2b	0.031	(1.2, 1.4)

For the cancellation rate we explore the system's behaviour with three rates, varying around the rates derived from the data. As functional forms of the cancellation rate are rarely reported upon in the literature, we identified only one manuscript provides cancellation measures for multiple scheduling intervals (Whittle et al. 2008). They found a similar monotonic relationship for patient initiated as well as clinic initiated cancellations. The exponential function parameters for the cancellation rate of Whittle et al. (2008) are $\chi_{\max} = 0.24$, $\chi_0 = 0.09$, and $C = 10$, which can be approximated with $\alpha = 0.05$, and is included in the experiments as well. Since some studies include cancellations in the no-show rates, we should be careful with the comparison of the various rates derived from these studies with our data-driven rates. However, they are valuable for analysis, since cancelled appointments may end up as empty appointment slots, and therefore reflecting no-show behaviour.

No study reported cancellation timing measures. Therefore, we base the timing behaviour on the

observations in the data analysis. In the analytical model we use an exponential distribution to determine the cancellation timing, whereas in the simulation, the cancellation rates from Institutions 1 and 2 have an empirically distributed scheduling interval dependent timing distribution based on the observations in Section 3.

Besides analysing the impact of the no-show and cancellation rate, we also analyse the impact of the arrival rate on the clinic behaviour. In line with Liu (2016), we expect higher arrival rates to result in lower booking horizons, and vice versa. Third, we consider multiple combinations of the cost coefficient weights θ_C and θ_B, to analyse the effect for various system settings. Fourth, we perform three case study experiments, with data from Institutions 1 and 2, to analyse the performance of our model on real life data, and to assess if the model is generalisable in practice. The two case studies of Institution 2 use the corresponding no-show rates from Table 1, and a cancellation rate of $\alpha = 0.093$ and $\alpha = 0.031$, as derived from Figures 2 and 3 respectively.

Considering the aforementioned parameters, we obtain a base case and 21 experiment instances. Table 3 gives an overview of the instances.

5.2. Experiment results

Table 4 provides an overview of the results of the queuing model experiments. Experiments 1–11 show the impact of the no-show and cancellation rates. For various no-show rates, an infinite booking horizon is optimal. These no-show rates have amongst the lowest asymptotes considered in the experiments, which supports the hypothesis that the lower the impact of no-shows, the longer the booking horizon can be. The impact of the cancellation rate to the optimal booking horizon is less clear. A small increase in booking horizon can be observed for lower cancellation rates, but no statistically significant difference is observed between the performance of the subsequent experiments. In additional experiments (not reported), we observe that low-traffic systems are more sensitive to no-show and cancellation behaviour of patients.

Experiments 12–15 evaluate the impact of the arrival rate. Table 4 shows a decrease in optimal booking horizon for higher values of λ. Thus, for high demand systems, it is beneficial to reduce the booking horizon, and possibly organising the clinic on a walk-in basis. This ensures that as many patients as possible can be served, as the patients that make an appointment, will most likely not end up as a no-show or cancellation. This corresponds to the finding of Liu (2016).

Experiments 16–19 evaluate the impact of various weights for the cost coefficients. We observe that when provider idle time is more important to the decision makers than rejections, the booking horizon is shorter than when idle time and rejections are equally valued. Therefore, the optimal booking horizon is dependent on the weights that decision makers assign to the cost coefficients, such as rejecting patients or provider idle time.

Table 4. Experiment results.

Exp no.	K^*	Days	Obj. value
Base case	21	3	3.430
1	∞	∞	3.701
2	13	2	2.995
3	∞	∞	4.851
4	∞	∞	4.218
5	∞	∞	4.464
6	∞	∞	4.420
7	∞	∞	4.540
8	20	3	3.288
9	21	3	3.374
10	22	3	3.468
11	23	3	3.568
12	∞	∞	3.507
13	13	2	2.692
14	7	1	1.693
15	7	1	−0.408
16	19	3	3.401
17	19	3	3.651
18	13	2	3.549
19	25	4	3.599
20	∞	∞	4.666
21	41	6	4.725

The case study experiments show that both for Institution 1 (exp. 5) and Institution 2a (exp. 20) an infinite booking horizon is optimal. For Institution 2b (exp. 21) a finite booking horizon of 41 slots (6 days) is optimal. Based on the data of Institution 2b, limiting the booking horizon to the proposed 6 days of the model, results in a cancellation and no-show fraction of 10.8% and 13.8% respectively, which is a reduction of 7.5% and 2.5% compared to limit on the booking horizon. This shows a clear advantage in reducing last-minute empty slots in practice by implementing a limited booking horizon.

In all experiments, the optimal booking horizon is found through a tradeoff between no-shows and cancellations, and patient rejections. For the base case, this is visualised in Figure 6. As expected, this figure shows that the no-show and cancellation probabilities increase with longer booking horizons, as patients are allowed to have longer waiting times. The rejection probability decreases with longer booking horizons, as more patients are admitted in the system.

6. Queueing model validation

The stylised queueing system is able to determine the optimal booking horizon for a clinic under restrictive assumptions. To assess the ability of the queuing system to capture reality, we need to evaluate the effects of these

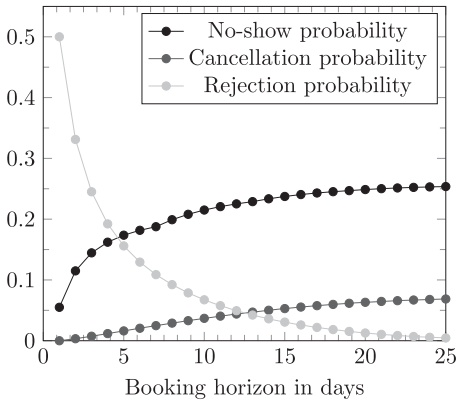


Figure 6. Average no-show, cancellation and rejection probabilities per booking horizon.

assumptions on the effectiveness of the optimal booking horizon. A first assumption in the queueing model was that the first patient in the queue is served. This implies that when a cancellation occurs, all patients in line after this cancelled appointment will be served one timeslot earlier. However, in practice empty spots due to cancellations are only filled when a new patient arrives that is willing to take that spot. Therefore, some slots might end up empty, if no patient arrives in the interval between the cancellation and the service of this specific appointment slot. A second assumption in the queuing model was that the cancellation rate is exponentially distributed with asymptote 1. However, the data analysis of Section 3 showed that the systems under consideration have lower asymptotes, and are bimodally distributed. Therefore, we need to analyse the impact of these assumptions on the system performance.

To validate our queuing model, we first compare our numerical results with empirical evidence, based on historical data. Because historical data on idleness and rejection performance is not available, we develop a data-calibrated simulation model to further assess the impact of the aforementioned assumptions in the queueing system and to evaluate the effectiveness of the optimal booking horizon results from the queuing model. The simulation model captures the bimodally distributed cancellation behaviour of the real system using the empirical distributions of Section 3.

6.1. Empirical validation based on historical data

To validate the results of the queueing model, we compare our modelled results with the no-show rates as derived from the data. As the queueing model does not allow for scheduling appointments further a maximum scheduling interval, we filter the data to only include appointments with scheduling intervals that are within this maximum scheduling interval. The queueing model results in no-show rates of 11.1, 9.3, and 8.4 percent respectively, which are similar to the real-life data no-show rates of 10.9, 9.3, and 8.7 percent.

6.2. Simulation-based validation

6.2.1. Simulation setup

The simulation model consists of a single server with a limited buffer of size $K-1$. The buffer represents the available appointment slots in the booking horizon, as derived from the queueing model of Section 4, where position 1 equals the first served slot, and position $K-1$ the last served slot. Together with the server, this makes the total number of positions in the system equal to K.

Patients arrive to the buffer according to a Poisson distribution with rate λ. Arriving patients are assigned the first available empty position in the buffer. If the buffer is full, patients are rejected.

When the deterministic server becomes empty, it processes the patient at position one in the buffer. If no patient is available at this position (independent of other possible patients in the queue), the server will remain empty for one timeslot. If a patient is available, and Δt equals the waiting time of this patient in the queue, with probability $\nu_{\Delta t}$ the patient is a no-show, and the server stays empty. With probability $1 - \nu_{\Delta t}$, the patient is seen, and is served. We assume deterministic service times with rate μ, equal to the daily capacity of the system, as we consider a tactical level appointment system design.

Patients may cancel their appointment when they are in the buffer. The cancellation probability depends on the patient's scheduling interval. The cancelled patient departs from the buffer, leaving an empty position in the buffer.

In the simulation model we measure several performance indicators. We record the proportion of rejected, cancelled, no-show, and seen patients, as well as the proportion of time the server is idle. This enables a comparison with the queuing system. Furthermore, we register the number of empty slots due to cancellations and an empty system.

We validated the simulation model by comparing the results of this model against the performance in practice. The no-show probabilities from the simulation are 10.9% (95%-CI: 10.7 –11.1), 9.3% (95%-CI: 9.2 –9.5), and 8.2%-(95%-CI: 7.9 –8.4), and cancellation probabilities 5.0% (95%-CI: 4.6 –5.4), 3.9% (95%-CI: 3.4 –4.4) and 3.7% (95%-CI: 3.4 –4.0) respectively, which are similar to the actual no-show rates of 10.9%, 9.3%, and 8.7% and cancellation rates of 4.9%, 4.0%, and 3.9% as derived from historical scheduling data with the same scheduling intervals. Therefore, the simulation model is considered valid.

The simulation model is developed in Tecnomatix Plant Simulation 11, and simulates 5 years, with a warm-up period of 75 days and 8 replications.

6.2.2. Simulation results

To evaluate whether the effects of neglecting the timing of cancellations has an impact on the analytical results, we simulated the system for each of the experiments with the corresponding K^* from Table 4. In the simulation the average percentage of idle time over all experiments was 24.9% (19.3% due to no-shows, and 5.5% due to an empty system). In the analytical results, the average idle time over all experiments was 25.7% (18.4% due to no-shows, and 7.3% due to an empty system).

The simulation shows that the number of empty slots in the queueing model is slightly overestimated (not significant, $p = 0.18$), as the system is 0.8% of the total time less idle on average. However, the idle time due to no-shows is significantly underestimated in the analytical experiments. Only simulation experiments 13–15 showed higher overall idle system probabilities compared to the analytical results. In these experiments, the system was overloaded with patients, which makes an empty system due to cancellations highly unlikely in the analytical model given the FCFS assumption. Therefore, the increase is primarily due to the impact of late cancellations. Note that the probability of an idle slot due to a late cancellation gets smaller when more patients arrive per time unit. The highest idle times in both the simulation and analytical model are seen in exp. 12, as there are often no patients in the system, since the average number of arrivals is lower than the capacity. The lowest idle times are seen in exp. 3, due to its low no-show rate.

Concluding, the outcomes of the stylised queueing system are valid, and therefore we consider them effective for strategic and tactical level decision making.

7. Discussion

No-show and cancellation behaviour of patients influence the performance of hospital's outpatient clinics, as less than 50% of all scheduled appointments may result in an actual patient being seen by the specialist. We investigated the scheduling interval in relation to no-show and cancellation rates, and found that an increasing scheduling interval results in higher no-show and cancellation probabilities. Therefore, clinics can benefit from limiting the possible scheduling intervals using a booking horizon, to minimise the effect of no-shows and cancellations. The optimal booking horizon is found through a tradeoff between the price of cancellations and no-shows and the price of rejection.

We developed an analytical queuing model to determine the optimal booking horizon, and provided a simulation study to evaluate the effectiveness of this model. Our results show that for systems with a high arrival rate, it is beneficial to limit the booking horizon. The impact of the no-show and cancellation rate showed to have a large impact on the optimal booking horizon in low-traffic systems. A limited booking horizon is also preferred for systems that highly value the utilisation of the providers. Note that for systems with an infinite booking horizon, it is still beneficial to schedule patients as early as possible, as this maximises the probability that the patient will show for the appointment.

In line with the current literature, we show that the no-show and cancellation rates are time-dependent. A longer scheduling interval results in higher no-show and cancellation probabilities. However, not only the occurrence of cancellations is related to the scheduling interval, but also the timing of cancellations. We are the first study to show that cancellation timing over multiple days follows a bimodal distribution, where peaks in cancellations are observed right after the creation of the appointment, and just before the actual appointment date. This corresponds with the literature that analysed reasons for cancellations, where scheduling conflicts, forgetting the appointment, and logistical challenges are frequently observed as main reasons for patient cancellations.

Our data-analysis and model provide insight into the impact of no-shows and cancellations. Where clinics tend to put more emphasis on reducing the number of no-shows compared to cancellations, this research showed that when focusing on the scheduling interval, the number of cancellations should get more attention, as the scheduling interval dependent no-show rate converges faster than the cancellation rate. Therefore, more efficiency gains can be derived in reducing the number of cancellations.

We showed the general applicability of our model by case studies of outpatient clinics of two hospitals in different health systems. We showed that efficiency gains can be achieved for certain combinations of no-show and cancellation rates derived from real-world scenarios, when a limited booking horizon is used. For low demand/low cancellation clinics it is optimal to have a long booking horizon in order to prevent unnecessary rejections, whereas for high demand/high cancellation clinics the optimal booking horizon is as short as possible.

Further research is required in the way how no-show education, reminders, and penalties impact the cancellation timing distribution, and the optimal booking horizon. We hypothesise that these interventions will cause more patients to cancel their appointment right before the actual appointment, which reduces the possibilities of reallocating their slots to new arrivals. In line

with this, immediate cancellations and late cancellations should be studied, ideally to be able to include these two cancellation types as individual rates to increase the validity of the model. Furthermore, literature has shown that new patients are more sensitive for long scheduling intervals than established patients (Davies et al. 2016). As very large datasets are required to define reliable time-dependent no-show and cancellation behaviour for subgroups, such as new and established patients, further research in large healthcare institutes with reliable data collection systems, is required to enable subgroup analyses. With the results of these analyses, the organisation of patient-specific appointment sequencing can be further investigated.

The benefit of using the queueing model over using the simulation model for a single parameter optimisation of the booking horizon is that it is an analytical method that requires few input data compared to setting up a computer simulation study. This makes the generic queuing model a valuable tool for strategic and tactical decision making. There are some restrictive assumptions that had to be made, such as that cancellations never lead to idle appointment slots, and for example exponential service times. The simulation model showed that despite these assumptions, the outcomes of the queueing model are effective for strategic and tactical level decision making. However, this model should not be used as an operational decision making tool, to for example analyse the individual patient's access times (although aggregate access time analyses can be easily performed). Operational level decisions, such as appointment scheduling and sequencing, require methods that are able to include more operational level details of the appointment scheduling process.

Although the simulation model already showed that the assumptions made in the queueing model are valid, future research can further extend the queueing model. For example, the service time in appointment systems starts on predefined timeslots, impacting the utilisation in case of empty appointment slots. This discrete-time nature of the appointment slots could be included in the queueing model to better capture the relationship between the booking horizon and scheduling interval (Creemers and Lambrecht 2010; Meisling 1958; Hernández-Díaz and Moreno 2009; Lozano and Moreno 2008).

Because we developed a generic queueing model, parameters such as the definition of cancellations and no-shows can easily be adapted to analyse other timescales. Typically, clinics that operate under shorter timescales, have more options to reuse cancelled slots on a short notice, compared to clinics with larger booking horizons. For example general practitioners would typically be able to reuse a cancelled appointment slot within the same day, whereas in most hospital outpatient clinics it is not possible to fill such cancelled slots the same day anymore. Our models can easily be adapted for these analyses, with data of clinics that use shorter time frames (see e.g. (Monahan and Fabbri 2018)). Note that scaling the model in time is important not only towards a smaller time frames, but also towards larger time frames depending on the type of clinic and its flexibility in filling slots as a response to cancellation behaviour.

Based on the presented approach and considerations, one of Institution 2's outpatient clinics has decided to limit their booking horizon to 8 weeks, as they were experiencing high cancellations and no-show behaviour in a highly utilised environment. Further research should analyse whether the predicted reductions in no-shows and cancellations are realised in practice, and what other considerations are of importance in booking horizon decision making. Further research in the implementation of short booking horizons is also required, as due to the asymptotic behaviour of no-shows and cancellations limited booking horizons (<3 months) are preferred, also for those patients that require appointments 6 or 12 months ahead of time. For example, patients may not be scheduled to follow up appointments that are further in the future than the optimal booking horizon. An alternative for patients that need (follow-up) appointments way ahead of time, is to maintain a call list. In such a system, a patient who was not given an appointment within the booking horizon, is added to this list and called to arrange an appointment once the booking horizon is extended. Another alternative is to implement a carefully designed admission control policy to reject patients. Our hospitals provide patients, who would otherwise be rejected due to fully booked calendars, an appointment slot in overtime within the booking horizon. Another policy could be to refer the patient to a partnering clinic. Each of these interventions can ensure that as many patients as possible are served, as patients scheduled within a shorter booking horizon are less likely to no-show or cancel.

No-show and cancellation behaviour not only influences the scheduling interval and booking horizon. Further research in incorporating these rates and the bimodal cancellation timing distribution in the design of (other elements of) appointment systems is required.

Acknowledgments

We thank Esra Sisikoglu Sir for her valuable contributions to the data analysis.

Disclosure statement

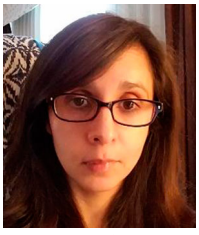
No potential conflict of interest was reported by the author(s).

Funding

This work was funded by the Netherlands Organization for Scientific Research (NWO)(Dutch Organization for Scientific Research), grant no. 406-14-128, and by the Mayo Clinic's Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery.

Notes on contributors

Gréanne Leeftink is an Assistant Professor in the Center of Healthcare Operations Improvement and Research (CHOIR) of the University of Twente, the Netherlands. Her research focuses on the design and optimization of integrated healthcare processes using Operations Management/Operations Research and Data Science techniques. She received her Ph.D. degree in Industrial Engineering and Management from the University of Twente in 2017, for which she was a researcher-in-residence at the University Medical Center Utrecht.



Gabriela Martinez is a Senior Decision Scientist at Siemens Healthineers. She received a Ph.D. in applied mathematics from Stevens Institute of Technology in 2011. Her research focuses on developing mathematical and simulation models to analyse the effectiveness of healthcare interventions considering the natural progression of disease and delivery of care.

Erwin W. Hans is a Full Professor Operations Management in Healthcare at the University of Twente in the Netherlands. He co-founded the Center of Healthcare Operations Improvement & Research (CHOIR, https://www.utwente.nl/en/choir/), the leading Netherlands' research center for OR/OM in healthcare. He works closely with several healthcare providers in the Netherlands, and has studied many applications in e.g. hospitals, rehabilitation and home care. He is an OR/OM lecturer on all academic levels and for healthcare professionals.

Mustafa Sir is a Senior Research Scientist at Amazon. Previously, he worked at Mayo Clinic focusing on developing clinical decision support systems using complex health data from sensors and electronic medical records. He holds a Ph.D. degree in Industrial and Operations Engineering from the University of Michigan in 2007.

Kalyan S. Pasupathy, Ph.D. is a faculty member in the Mayo College of Medicine and the Kern Center for the Science of Health Care Delivery. He is the Scientific Director for the Learning Laboratories and leads a research program in Information & Decision Engineering. Professor Pasupathy is an expert in systems science and health informatics and is focused on both, advancing the science and translating knowledge to improve care delivery demonstrated through his academic and practice leadership roles. He has over 20 years of experience leading and pioneering efforts, and conducting federally funded projects in designing and improving complex care delivery systems. He has received awards, created inventions, serves as a reviewer for several journals and for federal agencies, and is sought to consult or talk internationally.

ORCID

Gréanne Leeftink http://orcid.org/0000-0001-8835-5874

References

Ahmadi-Javid, Amir, Zahra Jalali, and Kenneth J Klassen. 2016. "Outpatient Appointment Systems in Healthcare: A Review of Optimization Studies." *European Journal of Operational Research* 258: 3–34.

Ancker, C. J., and A Gafarian. 1962. "Queueing with Impatient Customers Who Leave At Random." *Journal of Industrial Engineering* 13 (84-90): 171–172.

Bean, Andrew G, and James Talaga. 1992. "Appointment Breaking: Causes and Solutions." *Marketing Health Services* 12 (4): 14.

Blæhr, E. E., Rikke Søgaard, Thomas Kristensen, and Ulla Væggemose. 2016. "Observational Study Identifies Non-attendance Characteristics in Two Hospital Outpatient Clinics." *Danish Medical Journal* 63 (10

Centorrino, Franca, Miguel A Hernán, Giuseppa Drago-Ferrante, Melanie Rendall, Anthony Apicella, Gabriela Längar, and Ross J Baldessarini. 2001. "Factors Associated with Noncompliance with Psychiatric Outpatient Visits." *Psychiatric Services* 52 (3): 378–380.

Chariatte, Vincent, André Berchtold, Christina Akré, Pierre-André Michaud, and Joan-Carles Suris. 2008. "Missed Appointments in An Outpatient Clinic for Adolescents, An Approach to Predict the Risk of Missing." *Journal of Adolescent Health* 43 (1): 38–45.

Creemers, Stefan, and Marc Lambrecht. 2010. "Queueing Models for Appointment-driven Systems." *Annals of Operations Research* 178: 155–172.

Daggy, Joanne, Mark Lawley, Deanna Willis, Debra Thayer, Christopher Suelzer, Po-Ching DeLaurentis, Ayten Turkcan, Santanu Chakraborty, and Laura Sands. 2010. "Using No-show Modeling to Improve Clinic Performance." *Health Informatics Journal* 16 (4): 246–259.

Davies, Michael L, Rachel M Goffman, Jerrold H May, Robert J Monte, Keri L Rodriguez, Youxu C Tjader, and Dominic L Vargas. 2016. "Large-Scale No-Show Patterns and Distributions for Clinic Operational Research." In Healthcare, Vol. 15. Basel, Switzerland: Multidisciplinary Digital Publishing Institute

Denney, Joseph, Samuel Coyne, and Sohail Rafiqi. 2019. "Machine Learning Predictions of No-Show Appointments in a Primary Care Setting." *SMU Data Science Review* 2 (1): 2.

Foreman, D. M., and M Hanna. 2000. "How Long Can a Waiting List Be?" *The Psychiatrist* 24 (6): 211–213.

Gallucci, Gerard, Wayne Swartz, and Florence Hackerman. 2005. "Brief Reports: Impact of the Wait for An Initial

Appointment on the Rate of Kept Appointments At a Mental Health Center." *Psychiatric Services* 56: 344–346.
Green, Linda V, and Sergei Savin. 2008. "Reducing Delays for Medical Appointments: A Queueing Approach." *Operations Research* 56 (6): 1526–1538.
Guse, Clare E, Leanne Richardson, Mariann Carle, and Karin Schmidt. 2003. "The Effect of Exit-interview Patient Education on No-show Rates At a Family Practice Residency Clinic." *The Journal of the American Board of Family Practice* 16 (5): 399–404.
Hans, Erwin W, Mark Van Houdenhoven, and Peter J. H. Hulshof. 2012. "A framework for healthcare planning and control." In Handbook of healthcare system scheduling, 303–320. Boston, MA: Springer
Harris, Shannon LaToya. 2016. "Essays in appointment management." PhD diss., University of Pittsburgh.
Hawker, David S. J.. 2007. "Increasing Initial Attendance At Mental Health out-patient Clinics: Opt-in Systems and Other Interventions." *The Psychiatrist* 31 (5): 179–182.
Hernández-Díaz, A. G., and P. Moreno. 2009. "A Discrete-time Single-server Queueing System with An N-policy, An Early Setup and a Generalization of the Bernoulli Feedback." *Mathematical and Computer Modelling* 49 (5): 977–990.
Liu, Nan. 2016. "Optimal Choice for Appointment Scheduling Window Under Patient No-Show Behavior." *Production and Operations Management* 25 (1): 128–142.
Lozano, Macarena, and Pilar Moreno. 2008. "A Discrete Time Single-server Queue with Balking: Economic Applications." *Applied Economics* 40 (6): 735–748.
Meisling, Torben. 1958. "Discrete-Time Queuing Theory." *Operations Research* 6 (1): 96–105.
Mohammadi, Iman, Huanmei Wu Ayten Turkcan, Tammy Toscos, and Bradley N Doebbeling. 2018. "Data Analytics and Modeling for Appointment No-show in Community Health Centers." *Journal of Primary Care & Community Health* 9: 2150132718811692.
Monahan, Ken, and Daniel Fabbri. 2018. "Schedule-based Metrics for the Evaluation of Clinic Performance and Potential Recovery of Cancelled Appointments." *International Journal of Medical Informatics* 109: 49–54.
Moore, Charity G, Patricia Wilson-Witherspoon, and Janice C Probst. 2001. "Time and Money: Effects of No-shows At a Family Practice Residency Clinic." *Family Medicine-Kansas City-* 33 (7): 522–527.
Norris, John B, Chetan Kumar, Suresh Chand, Herbert Moskowitz, Steve A Shade, and Deanna R Willis. 2014. "An Empirical Investigation Into Factors Affecting Patient Cancellations and No-shows At Outpatient Clinics." *Decision Support Systems* 57: 428–443.
Partin, Melissa R, Amy Gravely, Ziad F Gellad, Sean Nugent, James F Burgess, Aasma Shaukat, and David B Nelson. 2016. "Factors Associated with Missed and Cancelled Colonoscopy Appointments At Veterans Health Administration Facilities." *Clinical Gastroenterology and Hepatology* 14 (2): 259–267.
Robinson, Lawrence W, and Rachel R Chen. 2010. "A Comparison of Traditional and Open-access Policies for Appointment Scheduling." *M&SOMt* 12 (2): 330–346.
Shah, Sachin J, Patrick Cronin, Clemens S Hong, Andrew S Hwang, Jeffrey M Ashburner, Benjamin I Bearnot, Calvin A Richardson, Blair W Fosburgh, and Alexandra B Kimball. 2016. "Targeted Reminder Phone Calls to Patients At High Risk of No-Show for Primary Care Appointment: A Randomized Trial." *Journal of General Internal Medicine* 31: 1460–1466.
Wang, Wen-Ya, and Diwakar Gupta. 2011. "Adaptive Appointment Systems with Patient Preferences." *Manufacturing & Service Operations Management* 13 (3): 373–389.
Whittle, Jeff, Gordon Schectman, Na Lu Bill Baar, and Michael F Mayo-Smith. 2008. "Relationship of Scheduling Interval to Missed and Cancelled Clinic Appointments." *The Journal of Ambulatory Care Management* 31 (4): 290–302.
Zacharias, Christos, and Michael Pinedo. 2014. "Appointment Scheduling with No-Shows and Overbooking." *Production and Operations Management* 23 (5): 788–801.

Appendix 1. Structural properties of the revenue function

This appendix provides the structural properties and its analytical results from the scheduling booking horizon problem as presented in Section 4.

A.1 Structural properties

Expanding the terms in (3), the revenue function can be expressed as $R(K) = \lambda T(K) - \lambda\theta_C$ with:

$$T(K) = \frac{\rho}{\delta}(1 - p_0(K))(1 + \theta_C) + p_K(K)(\theta_C - \theta_B) - P_N(K).$$

Given the monotone and asymptotic behaviour of the no-show rates and the long-run probabilities, it can be observed from the equation above that the existence of a finite $K \in \mathbb{Z}^+$ solution of (4) depends on the decrease rate of (1)–(2). Furthermore, we can solve the problem (4) by truncating the solution domain because $T(K)$ has a horizontal asymptote.

In order to gain some insights of the structure of (3), we will consider the particular case $\nu_j = \nu, j \in \mathbb{Z}^+$. In this case, the function $T(K)$ has a simple form:

$$T(K) = \frac{\rho}{\delta}(1 + \theta_C - \nu)(1 - p_0(K)) + p_K(K)(\theta_C - \theta_B).$$

Notice that if $\theta_B \geq \theta_c \geq 0$ then $T(K)$ is increasing in $\mathbb{Z}^+$ since it is expressed as the sum of increasing functions. Therefore, the booking horizon of the system can be as large as possible if the probabilities of no-shows behave relatively constant with respect to the capacity of the queue, and there is a preference to set up a higher penalty for blocking patients regardless of the values of ρ and δ. If $0 \leq \theta_B < \theta_C$ and let $w = \rho(1 + \theta_C - \nu)/\delta(\theta_C - \theta_B)$, then $T(K)$ can be expressed as follows:

$$T(K) = (\theta_C - \theta_B)\big(w + p_K(K) - wp_0(K)\big).$$

Let P_0 be the limit of $p_0(K)$ when $K \to \infty$, then from the equation above it follows that the queue capacity could be infinite if:

$$\sup_{K \in \mathbb{Z}^+} \big(p_K(K) - wp_0(K)\big) = -wP_0,$$

therefore, the idle penalisation weight w determines the existence of a finite queue capacity. The following algorithm could be used to obtain an upper bound of w for which an optimal finite queue capacity can be found:

A0 Set tolerance $\epsilon > 0$; $J = 1$; $K = 1$.
A1 Domain Truncation: while $p_0(J) - p_0(J + 1) > \epsilon$ then $J = J + 1$.

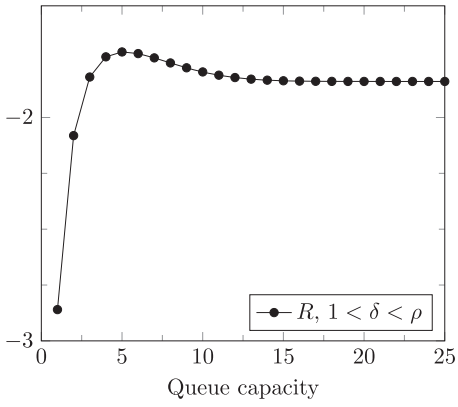


Figure A1. Revenue function with $\theta_C = 1.15, \theta_B = 1.0, K^* = 5$

A2 Idle weights: $w_i = p_K(i)/(p_0(k) - p_0(J))$, $i \leq J-1$.
A3 Upper bound: while $\text{argmax}_{j\leq J}(p_K(j) - w_K p_0(j)) \neq J$ then $K = K+1$.

Let w_K be the upper bound found by the algorithm with tolerance ϵ. A finite capacity could be found if $w \leq w_K$ which defines the following condition:

$$1 - \nu \leq \left(\frac{\delta}{\rho} w_K - 1\right)\theta_C - \frac{\delta}{\rho} w_K \theta_B. \tag{A1}$$

For example, applying the algorithm with $\epsilon = 1e-8$ we obtained $J = 24$, $K = 20$, $w_K = 104.08$ for a system with parameters $\mu = 3$, $\lambda = 6$, $\alpha = 0.6$, $\nu = 0.43$. Using (A1), we could select $\theta_c = 1.15, \theta_B = 1.0$ to obtain a queue capacity $K^* = 5$ (Figure A1). An optimal queue capacity $K^* = 6$ is obtained with $\theta_c = 4.0, \theta = 3.0$; (A1) is not satisfied if $\theta_B = 0, \theta_C = 0.01$, thus $K^* = \infty$.

The algorithm could be adapted by incorporating the asymptote of $P_N(K)$ in the condition of Step A1 when no-show probabilities are not constant. More details are provided in Appendix A.2.

A.2 Analytical results

This subsection provides the analytical results of the structural properties. In this section the notation $q_K(\cdot)$ is used to represent, as a function of queue length, the steady-state probability of rejection, i.e. the length of the queue is at full capacity. Let us notice that the functions $p_0(K)$ and $q_K(K)$ are defined in $\mathbb{Z}^+$, whereas the function $p_j(K)$ is defined in $\mathbb{Z}_j^+ := \{K \in \mathbb{Z}^+ | K \geq j\}$.

The monotonic properties of the steady-state probabilities with respect to length of the queue K (which represents the booking horizon) are summarised below.

Lemma A.1: *Given $K_1, K_2 \in \mathbb{Z}^+$ such that $K_1 \leq K_2$ then $q_K(K_1) \geq q_K(K_2)$, and $p_j(K_1) \geq p_j(K_2)$ for $j \in \mathbb{Z}^+$.*

Proof: Let $K_1, K_2 \in \mathbb{Z}^+$ such that $K_1 \leq K_2$, then:

$$\sum_{j=0}^{K_1} \Gamma(\delta)\frac{\rho^j}{\Gamma(\delta+j)} \leq \sum_{j=0}^{K_2} \Gamma(\delta)\frac{\rho^j}{\Gamma(\delta+j)},$$

since the summation involves non-negative numbers. It follows from the steady-state probabilities that $p_0(K_1) \geq p_0(K_2)$.

The recursive steady-state equations show that, for $j \in \mathbb{Z}^+$, the function $p_j(K)$ is non-increasing in its respective domain. Finally, we will show that the probability of rejection is non-increasing. Let $K \in \mathbb{Z}^+$, then using the closed-form equations we have:

$$q_K(K) - q_K(K+1) = \frac{\rho^K}{\Pi_{i=0}^{K-1}(\delta+i)} \times \left(p_0(K) - \frac{\rho}{\delta+K} p_0(K+1)\right)$$

$$q_K(K) - q_K(K+1) = \frac{\rho^K}{\Pi_{i=0}^{K-1}(\delta+i)} p_0(K)p_0(K+1) \times \left(1 + \sum_{j=0}^{K-1} \frac{\rho^j}{\Pi_{i=0}^{j-1}(\delta+i)} \left[\frac{1}{\delta+j} - \frac{1}{\delta+K}\right]\right) \geq 0$$

■

Let $\gamma(x,a)$ be the normalised lower incomplete gamma function defined as:

$$\gamma(x,a) = \frac{1}{\Gamma(a)} \int_0^x e^{-t} t^{a-1} dt.$$

Using the function $\gamma(\cdot,\cdot)$ we can reformulate the closed-form equations as (Ancker and Gafarian 1962):

$$p_0(K) = \left[1 + e^{\rho}\rho^{1-\delta}\Gamma(\delta)(\gamma(\rho,\delta) - \gamma(\rho,\delta+K))\right]^{-1}.$$

It follows from the expression above and Lemma A.1 that:

$$\lim_{K\to+\infty} p_0(K) = P_0 = \left[1 + e^{\rho}\rho^{1-\delta}\Gamma(\delta)\gamma(\rho,\delta)\right]^{-1}, \tag{A2}$$

$$\lim_{K\to+\infty} q_K(K) = 0, \tag{A3}$$

$$\lim_{K\to+\infty} p_j(K) = \frac{\rho^j\Gamma(\delta)}{\Gamma(\delta+j)} P_0, \quad j \in \mathbb{Z}^+. \tag{A4}$$

The limits shown in (A2)–(A4) result from that the gamma function grows faster than any power function.

Lemma A.2: *The rejection probability $\{q_K(K)\}_{K\in\mathbb{Z}^+}$ and $\{p_j(K)\}_{K\in\mathbb{Z}_j^+}$, for $j \in \mathbb{Z}^+ \cup \{0\}$, are convex sequences.*

Proof: A sequence is convex if its first difference is non-decreasing. Let $j = 0$, define the first-difference sequence $\{m_k\}_{k\in\mathbb{Z}^+}$ as:

$$m_k = p_0(K+1) - p_0(K) = -\frac{\rho^{K+1}\Gamma(\delta)}{\Gamma(\delta+K+1)} p_0(K)p_0(K+1).$$

We need to show that $m_k \leq m_{k+1}$, i.e, $m_k - m_{k+1} \leq 0$:

$$m_k - m_{k+1} = \frac{\rho^{K+1}\Gamma(\delta)}{\Gamma(\delta+K+1)} p_0(K+1) \times \left(\frac{\rho}{\delta+K+1} p_0(K+2) - p_0(K)\right). \tag{A5}$$

By Lemma A.1 we know that $q_K(\cdot)$ is non-increasing therefore:

$$\frac{\rho}{\delta+K}p_0(K+1)-p_0(K)\leq 0. \tag{A6}$$

Using (A6) in (A5) we have:

$$\begin{aligned} m_k-m_{k+1} &\leq \frac{\rho^{K+1}\Gamma(\delta)}{\Gamma(\delta+K+1)}p_0(K+1)\left(p_0(K+1)-p_0(K)\right)\\ &\leq 0, \end{aligned}$$

hence $\{p_0(K)\}_{K\in\mathbb{Z}^+}$ is a convex sequence. The convexity of $\{p_j(K)\}_{K\in\mathbb{Z}_j^+}$ follows from the steady-state equations.

Finally, let η_k be the first difference of the rejection probability sequence:

$$\begin{aligned} \eta_k &= q_K(K+1)-q_K(K)\\ &= \frac{\rho^K\Gamma(\delta)}{\Gamma(\delta+K)}\left(\frac{\rho}{\delta+K}p_0(K+1)-p_0(K)\right), \end{aligned} \tag{A7}$$

then:

$$\begin{aligned} \eta_k-\eta_{k+1} &= \frac{\rho^K\Gamma(\delta)}{\Gamma(\delta+K)}\left[\frac{2\rho}{\delta+K}p_0(K+1)-p_0(K)\right.\\ &\quad\left.-\frac{\rho^2}{(\delta+K)(\delta+K+1)}p_0(K+2)\right] \end{aligned} \tag{A8}$$

using the the closed-form of $p_0(\cdot)$ in equation (A8) we get, after algebraic manipulations, that $\eta_k-\eta_{k+1}\leq 0$, which shows the convexity of $\{q_K(K)\}_{K\in\mathbb{Z}^+}$. ■

Expanding the term in our objective function, the revenue function $R(K)$ can be expressed as $R(K)=\lambda T(K)-\lambda\theta_C$ with:

$$\begin{aligned} T(K) &= \frac{\mu}{\lambda}(1-p_0(K))(1+\theta_C)+P_N(K)\left(\frac{\mu}{\lambda}\theta_N-1\right)\\ &\quad+q_K(K)(\theta_C-\theta_B), \end{aligned} \tag{A9}$$

where θ_N represents a soft reward of the system when no-shows are observed. As mentioned before, the no-show probabilities of the system are described by a sequence $\{\nu_j\}_{j\in\mathbb{Z}^+}$ such that $\nu_j\leq\nu_{j+1}$ for all $j\in\mathbb{Z}^+$ and $\lim_{j\to+\infty}\nu_j=\nu^*$, $\nu^*\in[0,1]$. Then, $P_N(K)$ is bounded for all $K\in\mathbb{Z}^+$ and $\lim_{K\to+\infty}P_N(K)\leq\frac{\mu}{\lambda}(1-P_0)\nu^*$, because from the closed-form equations we have:

$$\begin{aligned} P_N(K) &= \sum_{j=0}^{K-1}p_j(K)\beta_j\nu_j=\frac{\mu}{\lambda}\sum_{j=1}^{K}p_j(K)\nu_{j-1}\\ &\leq \frac{\mu}{\lambda}(1-p_0(K))\nu_{K-1}, K\in\mathbb{Z}^+. \end{aligned} \tag{A10}$$

In order to gain some insights of the structure of the problem, we will consider the particular case $\nu_j=\nu, j\in\mathbb{Z}^+$. In this case, the function $T(K)$ has a simple form:

$$\begin{aligned} T(K)=T_\nu(K) &= \frac{\mu}{\lambda}(1-p_0(K))\left(1+\theta_C+\left(\frac{\mu}{\lambda}\theta_N-1\right)\nu\right)\\ &\quad+q_K(K)(\theta_C-\theta_B). \end{aligned} \tag{A11}$$

Lemma A.3: *If $\theta_B\geq\theta_C\geq 0$ then $T_\nu(K)$ is increasing in the domain $\mathbb{Z}^+$.*

Proof: Let $K_1,K_2\in\mathbb{Z}^+$ such that $K_1\leq K_2$, then:

$$\begin{aligned} T_\nu(K_1)-T_\nu(K_2) &= \frac{\mu}{\lambda}\left(1+\theta_C+\left(\frac{\mu}{\lambda}\theta_N-1\right)\nu\right)\\ &\quad\times(p_0(K_2)-p_0(K_1))\\ &\quad+(\theta_C-\theta_B)(q_K(K_1)-q_K(K_2)),\\ T_\nu(K_1)-T_\nu(K_2) &\leq (\theta_C-\theta_B)(q_K(K_1)-q_K(K_2)),\\ T_\nu(K_1)-T_\nu(K_2) &\leq 0, \end{aligned}$$

where the term $(1+\theta_C+(\frac{\mu}{\lambda}\theta_N-1)\nu)\geq 0$ since $0\leq\theta_N<1$. The first inequality follows from the decreasing property of $p_0(K)$. The last inequality is obtained from Lemma A.1 and the condition $\theta_B\geq\theta_C$. ■

An implication of Lemma A.3 is that the function $T_\nu(K)$ does not have a maximum in $\mathbb{Z}^+$ since

$$\begin{aligned} \sup_{K\in\mathbb{Z}^+}T_\nu(K) &= \lim_{k\to+\infty}T_\nu(K)\\ &= \frac{\mu}{\lambda}(1-P_0)\left(1+\theta_C+(\frac{\mu}{\lambda}\theta_N-1)\nu\right). \end{aligned}$$

Therefore, the booking horizon of the system can be as large as possible if the probabilities of no-shows behave relatively constant with respect to the capacity of the queue, and there is a preference to set up a higher penalty for blocking patients.

Another insight of Lemma A.3 is that if the function $T_\nu(K)$ has a maximum in $\mathbb{Z}^+$, then:

$$\max_{K\in\mathbb{Z}^+}T_\nu(K)>T_\nu^*,$$

where $T_\nu^*=\frac{\mu}{\lambda}(1-P_0)(1+\theta_C+(\frac{\mu}{\lambda}\theta_N-1)\nu)$.

Consequently, if $0<\theta_B<\theta_C$ we can truncate the domain of $T_\nu(K)$ by selecting a small tolerance number $\tau>0$ to find the smallest $\bar{K}\in\mathbb{Z}^+$ such that $p_0(K)-P_0<\epsilon$ and $q_K(K)<\epsilon$ for all $K\geq\bar{K}$, where $\epsilon=\tau/2((1+\theta_C+(\frac{\mu}{\lambda}\theta_N-1)\nu)+\theta_C-\theta_B)$. Then, the following optimisation problem always has a solution, and it can be solved by enumeration:

$$\max_{K\in\{1,\ldots,\bar{K}\}}T_\nu(K). \tag{A12}$$

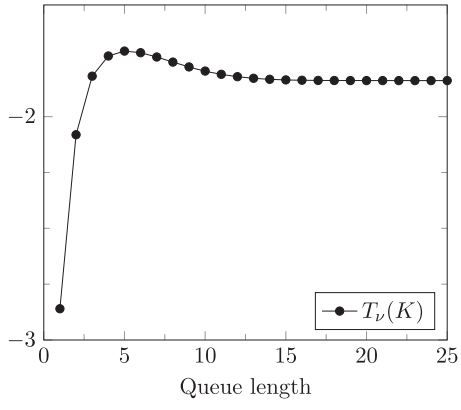


Figure A2. Revenue function with $\theta_C=1.15$, $\theta_B=1.0$, $\theta_N=0$, $K^*=5$

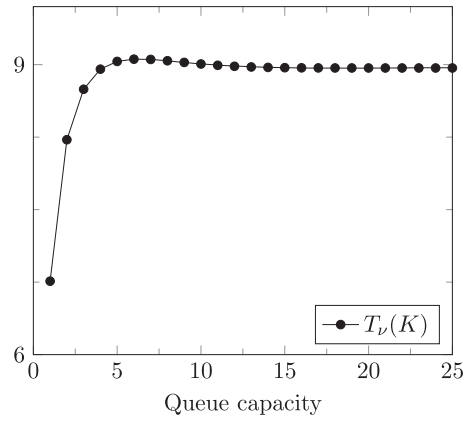


Figure A3. Revenue function with $\theta_C = 4$, $\theta_B = 3$,$\theta_N = 0$, $K^* = 6$

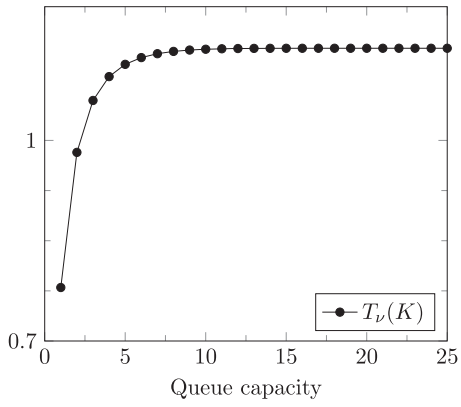


Figure A4. Revenue function with $\theta_C = 0.01$, $\theta_B = 0.0$, $\theta_N = 0$, $K^* = \infty$

Let us notice that, if $\bar{K}$ is a solution of (A12) then $T_\nu(K)$ does not have a maximum in $\mathbb{Z}^+$, because:

$$|T_\nu(\bar{K}) - T_\nu(K)| \leq |T_\nu(\bar{K}) - T_\nu^*| + |T_\nu(K) - T_\nu^*| < \tau,$$
$$\text{for all } K \geq \bar{K}.$$

In addition, by Lemma A.2, problem (A12) is a difference of convex (DC) optimisation problem. Therefore, the existence of a solution of (A12) such that $K < \bar{K}$, depends on the decrease rate of the functions $p_0(\cdot)$, $q_K(\cdot)$ and the magnitude of the penalisation and reward parameters as show in section III-B. For example, the Figures A2, A3, and A4 show revenue functions for a system with $\mu = 3$, $\lambda = 6$, $\alpha = 0,6$, $\nu = 0.43$, K^* indicates the value of the optimal queue capacity.

Finally, for a general form of $P_N(K)$ we still can solve the problem by enumeration as in (A12), because $P_N(K)$ has a horizontal asymptote. In addition, by (A10) the function $T(K)$ is dominated by a function that behaves like $T_\nu(K)$, therefore the following algorithm could be used to derive a condition between the queue and cost parameters:

A0 Set tolerance $\epsilon > 0$; $J = 1$; $K = 1$.
A1 Queue Capacity Truncation: Set $\tilde{J} = 1$ and while $p_0(\tilde{J}) - p_0(\tilde{J}+1) > \epsilon$ then $\tilde{J} = \tilde{J} + 1$.
A2 No-show Truncation: Set $\bar{J} = 1$ and while $\nu^{\bar{J}} - \nu^{\bar{J}+1} > \epsilon$ then $\bar{J} = \bar{J} + 1$.
A3 Set $J = \min\{\tilde{J}, \bar{J}\}$
A4 Idle weights: $w_i = p_K(i)/(p_0(k) - p_0(J))$, $i \leq J - 1$.
A5 Idle-weigh upper bound: while $\text{argmax}_{j \leq J}(p_K(j) - w_K p_0(j)) \neq J$ then $K = K + 1$.
A6 Approximate queue system bound:

$$\frac{\delta}{\rho}\left(1 + \theta_C + \left(\frac{\delta}{\rho}\theta_N - 1\right)\nu^K\right) \leq w_K(\theta_C - \theta_B)$$