Modeling Indirect Waiting Times with an M/D/1/K/N Queue

Anne Zander[a]

[a]*Institute of Operations Research, Karlsruhe Service Research Institute, Karlsruhe Institute of Technology, Karlsruhe, Germany*

Abstract

Indirect waiting times or access times of patients are an important indicator for the quality of care of a physician. Indirect waiting times are influenced by the panel size, i.e., the number of patients regularly visiting the physician. To study the nature of this influence we develop an M/D/1/K/N queueing model where we include no-shows and rescheduling. In contrast to previous work, we assume that panel patients do not make new appointments if they are already waiting. For a given panel size we calculate the steady state probabilities for the indirect queue length and further aspects such as the effective arrival rate of patients. We compare those results to the outcomes of a simulation and show that the simplifications we used in the analytical model are verified. The queueing model can help physicians to decide on a panel size threshold in order to maintain a predefined service level with respect to indirect waiting times.

Keywords: Health Services, Panel Size, Traditional Appointment Policy, Access Time, No-shows, Queueing Model

1. Introduction

Appointment planning matches patient demand and health care provider supply. Good reviews of this research area can be found in (Cayirli and Veral,

Email address: anne.zander@kit.edu (Anne Zander)

2009) and (Gupta and Denton, 2008). A more recent review of online appointment planning can be found in Chapter 2 of the PhD thesis (Braaksma, 2015).
Patients can stay informed about health care providers for example through evaluation portals. Hence, there is an incentive for doctors to pay more attention to service aspects such as waiting times. In this paper, we focus on the indirect waiting time or access time, defined as the elapsed time between the moment the appointment is made and the actual appointment time. Long indirect waiting times can lead to a deterioration in patients' health. Some studies also show that they may increase the probability of a patient being a no-show (Gallucci et al., 2005). Consequently, doctors should reduce their indirect waiting times in order to deliver better service to patients and avoid idle time.
For our investigation, we assume that each doctor has a panel, i.e., a group of patients who visit on a regular basis. We also assume that all the demand comes from that panel. The doctor could fix an indirect waiting time service level; e.g., on average, patients should not wait more than two weeks for treatment. Then, the doctor must manage the size of the panel in order to achieve this service level.
Our aim is to model the scheduled appointment queue in order to relate panel sizes to indirect waiting times.

2. Literature

The model presented in this paper is based on one of the models of (Green and Savin, 2008). They present two queuing models (M/D/1/K and M/M/1/K) in order to link the panel size with the average indirect queue length. They assume that appointment requests are only coming from the panel and that they come with a constant rate which is independent of the indirect queue length. Their aim is to find panel sizes which allow the doctor to implement an open access policy where patients can only make appointments for the same day. They presume that an open access system can be installed if the expected probability of getting a same day appointment for a patient is above a certain threshold, e.g., 80%. This means that 80% of the time the indirect queue length is shorter than a day. Given the threshold, an upper bound for the panel size can then be determined.
In (Liu and Ziya, 2014) they decide about the panel size and the offered

capacity in order to maximize profit. There is a fixed reward for treating a patient and they assume costs for overtime.
In (Zacharias and Armony, 2013) both direct and indirect waiting times are considered. Again decisions on panel sizes and capacities offered are made in order to maximize profit.
Further, we mention (Balasubramanian et al., 2010) and (Ozen and Balasubramanian, 2013) where patients are divided into groups representing different demands, e.g., average number of appointment requests per year. In a multi-provider clinic patients can then be relocated from one doctor panel to another in order to achieve a workload balance between the doctors such that a minimum number of patients have to change their doctor.
In contrast to (Green and Savin, 2008), we also want to investigate indirect queues of doctors that operate under a traditional appointment policy, i.e., every patient has to make an appointment in a given planning horizon. In general, the panel sizes under consideration can be bigger than those suitable for open access as waiting times greater than one day are allowed. But then a substantial part of the panel might be waiting in the indirect queue. Assuming that patients waiting do not make new appointments makes it necessary to make the demand rate dependent on the indirect queue length. Hence, our contribution is to extend the model of (Green and Savin, 2008) in order to include the analysis of traditional appointment systems.

3. Model

We present a queueing model and a simulation model which extend the M/D/1/K queueing model and the simulation model presented in (Green and Savin, 2008). We assume a single server queue modeling the appointment schedule and therefore the indirect waiting time. Here, we assume that appointment requests are only coming from the panel and that patients always accept the next available appointment. The difference to (Green and Savin, 2008) is that we assume that patients waiting in the queue will not make new appointments whereas in (Green and Savin, 2008) the rate of appointment requests is constant independent of the queue length. We will use the same notation as in (Green and Savin, 2008). We assume that the doctor can treat a fixed number of patients every day and therefore we assume deterministic service times of length T. By λ we denote the individual patient appointment request rate. We approximate the arrival process for every ser-

vice period as a Poisson process with a parameter dependent on the number of panel patients not waiting or getting treatment. Therefore, the panel size being N, we define $\alpha_i(k) = \frac{(\lambda(N-i)T)^k}{k!}e^{-\lambda(N-i)T}$ as the (approximate) probability that k patients arrive during a service period given that i patients are in the system (waiting or getting treatment). Further, we assume a finite queue capacity of K (corresponding to a finite booking horizon). Following the notation of (Osaki, 1992, p. 233) we denote our model as an M/D/1/K/N queue.

As in (Green and Savin, 2008) we will use a no-show function γ which gives the now-show probability of patients dependent on their indirect waiting time. For tractability reasons we calculate the no-show probability based on the queue length at the time of a patient's treatment rather than on his or her time of arrival. In addition, no-shows will schedule a new appointment with probability r.

With a similar argumentation as in (Green and Savin, 2008), we derive analytical expressions for the stationary distribution of the number of patients in the system, $\pi(k)$ being the probability that $k = 0, \ldots, K$ are in the system. We use $\rho = \lambda NT$.

Proposition 1. *The stationary distribution of the number of patients in the system is given by*

$$\begin{aligned}
\pi(0) =& \frac{1 - r\gamma(K)}{1 - r\gamma(K) + \rho(\sum_{i=0}^{K-1} f(i)) - r\sum_{i=1}^{K-1}(\gamma(K) - \gamma(i))f(i)} \\
\pi(k) =& \frac{(1 - r\gamma(K))f(k)\frac{N}{N-k}}{1 - r\gamma(K) + \rho(\sum_{i=0}^{K-1} f(i)) - r\sum_{i=1}^{K-1}(\gamma(K) - \gamma(i))f(i)}, \\
k =& 1, \ldots, K-1 \\
\pi(K) =& 1 - \frac{(1 - r\gamma(K))(\sum_{i=0}^{K-1} f(i)\frac{N}{N-k})}{1 - r\gamma(K) + \rho(\sum_{i=0}^{K-1} f(i)) - r\sum_{i=1}^{K-1}(\gamma(K) - \gamma(i))f(i)}
\end{aligned}$$

where $f(k)$ is a recursion with

$$\begin{aligned}
f(0) &= 1 \\
f(1) &= \frac{1}{(1 - r\gamma(0))\alpha_1(0)} - \frac{\alpha_0(0)}{\alpha_1(0)}
\end{aligned}$$

$$f(k+1) =$$
$$\frac{1}{(1-r\gamma(k))\alpha_{k+1}(0)}(f(k)-(1-r\gamma(k))\alpha_0(k)-r\gamma(k-1)\alpha_0(k-1))$$
$$-\frac{1}{(1-r\gamma(k))\alpha_{k+1}(0)}(\sum_{i=1}^{k}((1-r\gamma(k))\alpha_i(k+1-i)+r\gamma(k-1)\alpha_i(k-i))f(i)),$$
$$k=1,\ldots,K-2$$

We also build a simulation model again following (Green and Savin, 2008) in order to avoid the approximation we used employing the no-show function. Further, the arrival process can be approximated by a binomially distributed random number for every service period. This way we do not make an approximation error for cases when almost the whole population is waiting in the queue. Moreover, as in (Green and Savin, 2008) the assumption that every patient accepts the next free appointment can be relaxed.

4. Numerical Experiments

We use the same parameter settings (see Table 1) as in (Green and Savin, 2008). We assume 20 appointment slots per day.

Parameter	Definition	Value
K	Queue capacity	400
λ	Individual arrival rate	$\frac{0.008}{day}$
T	Service time	$0.05\,days$
r	Rescheduling probability	1
γ_0	Min no-show probability	0.01
γ_{max}	Max no-show probability	0.31
C	Sensitivity parameter	$50\,days$

Table 1: Parameter settings

We also use the same no-show function: $\gamma(k) = \gamma_{max} - (\gamma_{max} - \gamma_0)e^{-\lfloor kT \rfloor / C}$ where k is the queue length and $\gamma(k)$ the probability of being a no-show. In Figure 1 the average queue length dependent on the panel size is depicted for different models.

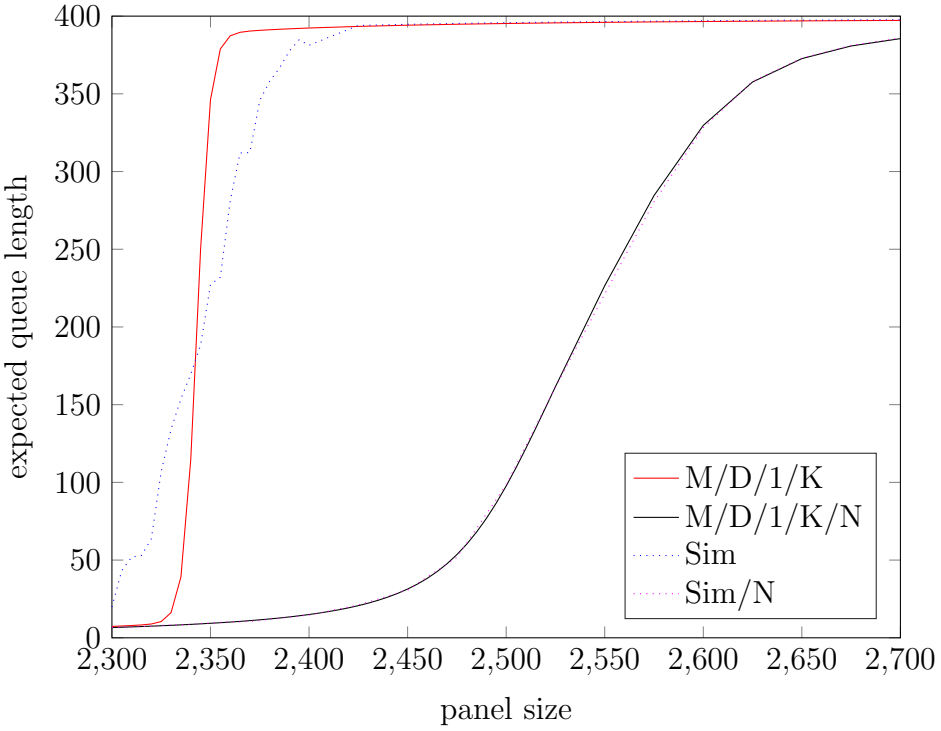

Figure 1: Expected queue length for different panel sizes for the two models and the corresponding simulations

The models M/D/1/K and Sim correspond to the models from (Green and Savin, 2008) whereas the models M/D/1/K/N and Sim/N correspond to the models presented in this paper. Both simulations started with an empty queue and were simulated for 40000 warm-up periods (corresponding to 7 years) and 10000 more periods to collect data. The implementation is based on (Koza, 2014). First, our results for the M/D/1/K model differ slightly from the results presented in (Green and Savin, 2008). The fundamental behavior of the curve is the same but the transition from an almost empty to an almost full schedule happens for bigger panel sizes. Comparing the different model approaches, the increase of the M/D/1/K curve is much steeper and happens for smaller panel sizes than the increase of the M/D/1/K/N curve. Further, we see that the difference between the results of the analytical model and the simulation are much smaller in our case.
It is also interesting to note that the queue length distributions are fundamentally different between M/D/1/K and M/D/1/K/N. In Figure 2 queue

length distributions with an expected queue length of circa 200 for the two models are depicted. For the two models this expected queue length is attained for different panel sizes.

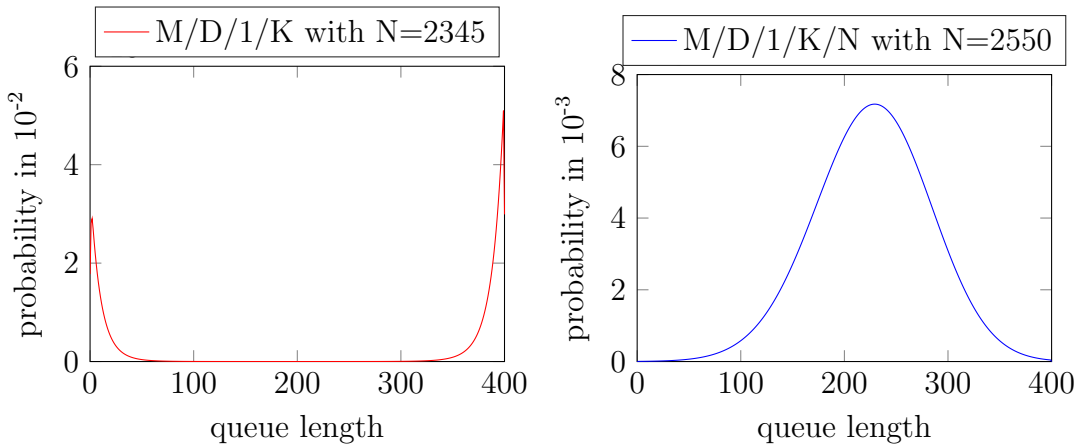

Figure 2: Queue length distributions in comparison

In our model the queue length will most likely be found around the expected value whereas for the M/D/1/K model the queue length oscillates between empty and full which is of course an unattractive behavior. It can be shown with the simulation that the initial queue length has a great influence on the queue length even after a lot of simulation periods because the queue either tends to a fully booked schedule or to an empty schedule and then stays there for a long time.

By rate we denote the request rate only generated by the patients not waiting or rescheduling. We assume that patients in the indirect queue do not make new appointments. That means that for long indirect queues patients in the queue do not make appointments but they would if they were not already waiting. We call this the hidden demand. Hence, the demand seems smaller than it should actually be. Conversely, due to the effect of rescheduling (also more prominent for long indirect queues) there is an extra demand in addition to the rate leading to a higher effective demand rate. Moreover, it is important to note that patients get rejected when the indirect queue is very long. In Table 2 we quantify those effects for different panel sizes. The calculations of the effects are based on the analytical model.

Effect/Panel size	2500	2600	2700	2800
Average rate	19.22	18.16	18.52	19.26
Average hidden demand rate	0.78	2.64	3.08	3.14
Average effective rate	19.97	20.04	20.63	21.41
Proportion of rejected patients	0.00	0.00	0.03	0.07
Proportion of no-shows	0.04	0.09	0.11	0.11

Table 2: Effects for different panel sizes, rates are measured per day

As you can see in Table 2, the average rate does not change much for large panel sizes but the effective rate increases due to the increasing proportion of no-shows that reschedule. In addition, the hidden demand rate increases substantially and the proportion of rejected patients is significant starting from a panel size of around 2600. This shows that a doctor should not only consider the average indirect waiting time but also the proportion of rejected patients and the hidden demand rate as those are indicators for a possible lack of care.

5. Conclusion and Outlook

We present a queueing model and a simulation model in order to connect panel sizes with the distribution of indirect waiting times. The model can help doctors operating under the traditional appointment policy to decide about a maximal panel size in order to achieve a service level with respect to their indirect waiting times and other effects such as the proportion of rejected patients and the hidden demand rate. Mathematically, the contribution of this paper is the analytical distribution of an M/D/1/K/N queue where the arrival rate is dependent on the queue length.
Possible future work on the model include a sensitivity analysis for the model parameters and numerical experiments considering measures such as seeing a given share of the patients in a fixed time period. Then, other patient behaviors such as balking, rescheduling directly (even if not being a no-show) if the queue is very long or leaving the panel because of long indirect waiting times could be integrated. Moreover, patients could book several appointments at a time. They may not always book the next available appointment. In general, physicians do not operate under a purely traditional appointment system (only patients with appointments are treated) but also allow for walk-ins, e.g., urgent cases. This option should be included in the model

considering that the longer the indirect queue the more likely patients will just walk-in. Furthermore, the idea of (Balasubramanian et al., 2010) and (Ozen and Balasubramanian, 2013) that patients belong to different demand groups could be integrated. Then, not only the panel size but also the case mix is relevant. In addition, the model could be extended to a vacation queueing system as in (Creemers and Lambrecht, 2009a) and (Creemers and Lambrecht, 2009b).

References

Balasubramanian, H., Banerjee, R., Denton, B., Naessens, J., and Stahl, J. (2010). Improving clinical access and continuity through physician panel redesign. *Journal of general internal medicine*, 25(10):1109–15.

Braaksma, A. (2015). *Timely and Efficient planning of TReatments through Intelligent Scheduling*. PhD thesis, University of Twente.

Cayirli, T. and Veral, E. (2009). Outpatient Scheduling in Health Care: a Review of Literature. *Production and Operations Management*, 12(4):519–549.

Creemers, S. and Lambrecht, M. (2009a). An advanced queueing model to analyze appointment-driven service systems. *Computers & Operations Research*, 36(10):2773–2785.

Creemers, S. and Lambrecht, M. (2009b). Queueing models for appointment-driven systems. *Annals of Operations Research*, 178(1):155–172.

Gallucci, G., Swartz, W., and Hackerman, F. (2005). Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. *Psychiatric services (Washington, D.C.)*, 56(3):344–6.

Green, L. V. and Savin, S. (2008). Reducing Delays for Medical Appointments: A Queueing Approach. *Operations Research*, 56(6):1526–1538.

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

Koza, D. F. (2014). Appointment Scheduling for Medical Practices. Master's thesis, Karlsruhe Institute of Technology.

Liu, N. and Ziya, S. (2014). Panel size and overbooking decisions for appointment-based services under patient no-shows. *Production and Operations Management*, 23(12):2209–2223.

Osaki, S. (1992). *Applied stochastic system modeling*. Springer, Berlin.

Ozen, A. and Balasubramanian, H. (2013). The impact of case mix on timely access to appointments in a primary care group practice. *IIE Transactions*, 16(2):101–18.

Zacharias, C. and Armony, M. (2013). Joint Panel Sizing and Appointment Scheduling in Outpatient Care. *Working paper 5.6. 1.*