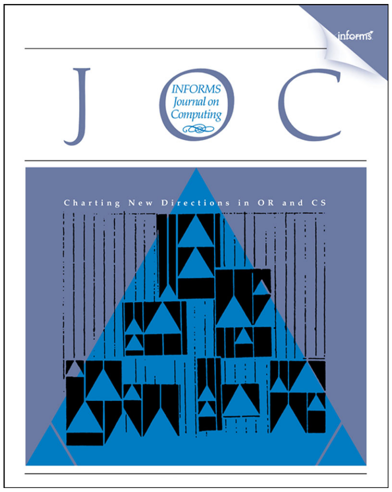


INFORMS Journal on Computing

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

A Data-Driven Model of an Appointment-Generated Arrival Process at an Outpatient Clinic

Song-Hee Kim, Ward Whitt, Won Chul Cha

http://pubsonline.informs.org/journal/ijoc/

INFORMS JOURNAL ON COMPUTING
Vol. 30, No. 1, Winter 2018, pp. 181–199
ISSN 1091-9856 (print), ISSN 1526-5528 (online)

A Data-Driven Model of an Appointment-Generated Arrival Process at an Outpatient Clinic

Song-Hee Kim,[a] **Ward Whitt,**[b] **Won Chul Cha**[c]

[a] Data Sciences and Operations, Marshall School of Business, University of Southern California, Los Angeles, California 90089; [b] Industrial Engineering and Operations Research, Columbia University, New York, New York 10027; [c] Department of Emergency Medicine, Samsung Medical Center, Seoul, Korea

Contact: songheek@marshall.usc.edu, http://orcid.org/0000-0002-3106-5726 (S-HK); ww2040@columbia.edu, http://orcid.org/0000-0003-4298-9964 (WW); docchaster@gmail.com, http://orcid.org/0000-0002-2778-2992 (WCC)

Received: June 20, 2016
Revised: January 2, 2017; June 27, 2017
Accepted: June 29, 2017
Published Online: February 5, 2018

https://doi.org/10.1287/ijoc.2017.0773

Abstract. We develop a high-fidelity simulation model of the patient arrival process to an endocrinology clinic by carefully examining appointment and arrival data from that clinic. The data include the time that the appointment was originally made as well as the time that the patient actually arrived, as well as if the patient did not arrive at all, in addition to the scheduled appointment time. We take a data-based approach, specifying the schedule for each day by its value at the end of the previous day. This data-based approach shows that the schedule for a given day evolves randomly over time. Indeed, in addition to three recognized sources of variability—(i) no-shows, (ii) extra unscheduled arrivals, and (iii) deviations in the actual arrival times from the scheduled times—we find that the primary source of variability in the arrival process is variability in the daily schedule itself. Even though service systems with arrivals by appointment can differ in many ways, we think that our data-based approach to modeling the clinic arrival process can be a guideline or template for constructing high-fidelity simulation models for other arrival processes generated by appointments.

History: Accepted by Bruno Tuffin, Area Editor for Simulation.
Funding: Support was received from NSF [Grants CMMI 1265070 and CMMI 1634133].
Supplemental Material: The online supplement is available at https://doi.org/10.1287/ijoc.2017.0773.

Keywords: simulation stochastic input modeling • simulating appointment-generated arrival processes • scheduled arrivals in service systems • outpatient clinics • data-driven modeling • stochastic models in healthcare

1. Introduction

In this paper we aim to contribute to simulation stochastic input modeling. In particular, we develop a data-based approach for creating high-fidelity stochastic models of arrival processes generated by appointments. We do that so that the arrival process model can be part of a full simulation model used to improve operations (e.g., to improve throughput, control individual workloads, set staffing levels, and allocate other resources), with the goal of efficiently providing good service in a service system with arrivals by appointment.

In particular, we create a data-driven stochastic arrival process model for a doctor in an outpatient clinic by carefully examining patient appointment and arrival data from an outpatient clinic. The data include the day and time of each appointment, when the appointment was originally made, an indicator for whether the scheduled arrival actually came, the time of arrival (if the patient came), and an indicator for cancellation (if the patient cancelled). All time stamps are measured to the nearest second.

1.1. A Long History of Modeling and Analyzing Outpatient Clinics

There is a long history of modeling and analyzing outpatient clinics and other healthcare systems, with notable early work by Bailey (1952), Welch and Bailey (1952), Fetter and Thompson (1965), and Swartzman (1970); surveys by Jun et al. (1999), Cayirli and Veral (2003), Jacobson et al. (2006), and Gupta and Denton (2008); and edited reviews by Hall (2006 and 2012). Depending on the studies' focus, the large literature can be divided roughly into three types of analyses.

The first type of analysis is a full analysis of an outpatient clinic to make operational improvements. As illustrated by the seminal paper by Fetter and Thompson (1965), outpatient clinics can be represented as a complex network of queues associated with the reception area, nurses, labs, and doctors. Patients often follow different paths through the clinic, depending on many factors, such as the doctor whom they are scheduled to see, their medical condition, and the results of medical tests. The system complexity has made simulation the dominant choice for detailed analysis of a clinic. Many successful simulation studies have been conducted, as can be seen from Swisher et al. (2001), Harper and Gamlin (2003), Guo et al. (2004), Chand et al. (2009), and Chakraborty et al. (2010).

The second type of analysis is designing an effective appointment system. Most outpatient clinics have a substantial portion of their arrivals scheduled in advance, i.e., generated by an appointment system. A large part of the literature is devoted to designing an effective appointment system, as can be seen from surveys by Cayirli and Veral (2003) and Gupta and Denton (2008) and other works by Liu et al. (2010), Luo et al. (2012), and Liu and Ziya (2014).

The third type of analysis is conducting a performance analysis of queueing models based on assumed properties of clinic arrival processes. It is recognized that appointment-generated arrival processes differ from arrival processes where customers independently decide when to arrive. In theory, appointment-generated arrival processes should have a nearly periodic structure determined by appointment time slots. However, studies have shown that arrival processes can be significantly variable because of patient no-shows, unscheduled patient arrivals, and patient earliness or lateness. Studies have found that no-show rates vary across different services and patient populations: the reported no-show rates are as low as 4.2% at a general practice outpatient clinic in the United Kingdom (Neal et al. 2001) and as high as 31% at a family practice clinic in South Carolina (Moore et al. 2001). Ever since the seminal papers of Bailey (Bailey 1952, Welch and Bailey 1952), studies have analyzed queueing models that reflect key structural properties of appointment-generated arrival processes, e.g., see Kaandorp and Koole (2007), Hassin and Mendel (2008), Araman and Glynn (2012), Jouini and Benjaafar (2012), Feldman et al. (2014), Honnappa et al. (2015), Wang et al. (2014), and Zacharias and Pinedo (2014).

1.2. Carefully Probing into One Clinic Arrival Process

In this paper, we do not follow any of the three time-tested approaches discussed above. We instead devote this entire paper to carefully examining arrival data from an outpatient clinic appointment system. In doing so, we aim to construct a high-fidelity stochastic arrival process model that can be part of a simulation model or analytic queueing model that can be used to improve the performance of the clinic. We want to understand the consequence of existing appointment schedules; we do not consider alternative scheduling algorithms.

The data were collected over a 13-week period from July 2013 to September 2013 from the endocrinology outpatient clinic of the Samsung Medical Center in South Korea. Sixteen doctors work in this clinic, but patients make an appointment to see a particular doctor, so each arriving patient knows which doctor he or she will meet. (The clinic is strict about having each patient see the scheduled doctor.) Hence, each doctor operates as a single-server system. Each doctor works within a subset of available days and shifts, with three shifts available: morning (A.M.) shifts, roughly from 8:30 A.M. to 12:30 P.M.; afternoon (P.M.) shifts, roughly from 12:30 P.M. to 4:30 P.M.; and full-day shifts. See Table 7 in the online supplement for the distribution of shifts for each doctor. The data include the day and time of each appointment and when the appointment was made. The data also have an indicator for whether the scheduled patient actually came and, if so, what was the time of arrival, and, if the patient did not come, if and when there was a cancellation. If the arrival did not come and there was no cancellation, the appointment is regarded as a no-show. All time stamps are measured to the nearest second.

We focus on patient arrivals during the A.M. shifts of one doctor to develop our data-driven approach. This doctor was selected from among the 16 candidate doctors because of his relatively high volume of patients: he worked for a total of 22 A.M. shifts (12 on Tuesdays and 10 on Fridays) and 22 P.M. shifts (11 on Mondays, two on Wednesdays, and nine on Thursdays) during our study period. We analyze the data in steps, leading up to a full stochastic process model. We do not immediately present the final model because we regard the process leading up to the model as more important than the resulting model for the arrival process.

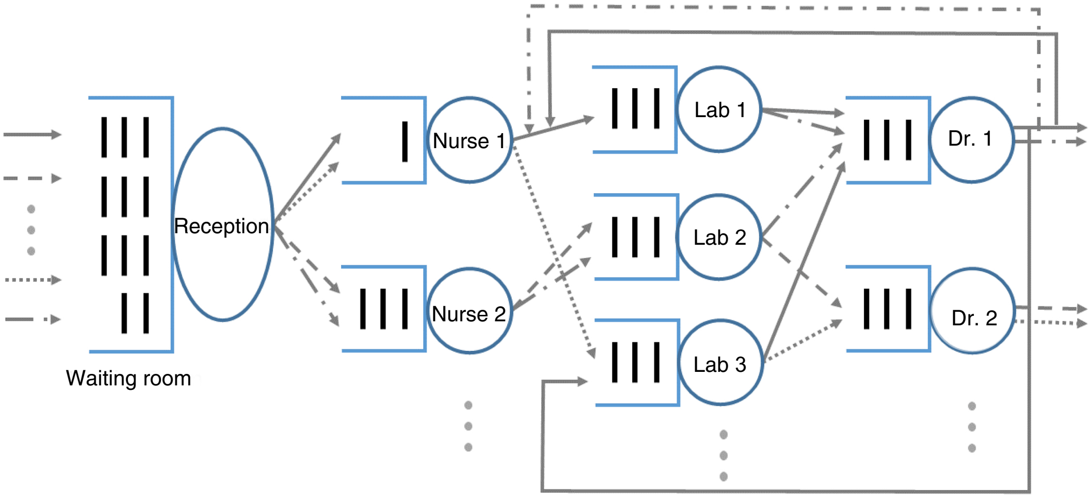
To confirm our approach for the clinic, in the online supplement we also carry out the entire analysis again for three other shifts. We consider the P.M. shifts of the same doctor to contrast A.M. and P.M. shifts, and we consider A.M. and P.M. shifts of other doctors.

1.3. The Clinic Viewed as an Open Network of Queues

Although we focus only on creating an arrival process model and do not analyze the clinic operations, understanding that how the clinic operates is important to appreciate the stochastic model we create. First, this clinic, just like most other medical clinics, does not operate as a simple conventional single-server queue, even though the patients have appointments with a designated doctor. A conventional single-server queue has customers (patients) arrive, wait, and then receive a single uninterrupted service by a server (the doctor), who is dedicated to them and is present with them for the entire service time. In contrast, in a medical clinic, a patient might spend an hour or two in the system after he starts service, while the doctor might spend only a few minutes with the patient. That is indeed what happens in the clinic we study.

Second, we envision the stochastic arrival process model we create as being part of a larger model of the entire clinic, as in previous studies mentioned in Section 1.1. In particular, we think of the endocrinology clinic with 16 doctors being modeled as a multiclass open network of queues, as depicted in Figure 1. The different colored arrows represent the flows of different classes of patients, e.g., classified by their medical

Figure 1. (Color online) The Clinic Viewed as a Multiclass Open Network of Queues



condition and the doctor they are scheduled to see. Figure 1 shows feedback flows, because patients might need medical procedures both before and after seeing the doctor. Since many of these patients use the same resources, we have the usual issues of resource sharing, which queueing network models are designed to address.

This modeling approach is consistent with the previous use of open queueing network models for complex manufacturing systems, as illustrated by Whitt (1983) and Segal and Whitt (1989). Such queueing network models can be used to design new clinics and to study possible changes to existing clinics. There are many questions the model can address: e.g., what would happen if the mix of doctor specialties changes? Or what would happen if the number of patients seen by each doctor on each shift changes? Or what happens if the punctuality can be improved? Just as with previous stochastic simulation models, this stochastic model makes it possible to answer various what-if studies, which is not possible—or at least not easy—just using data, and to assess the statistical precision of simulation estimates.

1.4. Organization of the Paper

In Section 2, we first examine the observed schedules to infer an underlying *master schedule*. The master schedule usually specifies the total number of appointment slots for each day, the length of each appointment slot, and the number of patients to be scheduled for each appointment slot (Liu et al. 2016). Then we view the (*actual*) *schedule* as a random modification of the master schedule. We find that the main deviation from a regular deterministic arrival pattern of a master schedule is variability in the schedule itself.

In Section 3, we view the patient arrivals as a random modification of the schedule and examine to what extent the arrivals adhere to the schedule. In Section 4, we study the pattern of arrivals over each day and directly compare the arrivals to the schedule. In Section 5, we provide mathematical representations of the stochastic counting processes for the scheduled and actual arrivals, as well as a simple parsimonious model that may be a convenient substitute for mathematical analysis. We provide a classification for appointment-generated arrival processes in Section 6, which provides a basis for comparing the different doctors in this clinic with each other and with doctors in other clinics. We conclude in Section 7. Together, Sections 6 and 7 provide an overview of the proposed modeling approach that we think is broadly applicable.

We also present supplementary material, including the analysis of three other doctor shifts, in the online supplement and in our longer study of all doctors in the clinic Kim et al. (2015a).

1.5. Main Contributions of the Paper

First, we provide a general framework for analyzing and modeling an appointment-generated arrival process given appointment scheduling and arrival data. There is an active ongoing effort to advance the understanding of and to improve the operations of outpatient clinics; e.g., see Zacharias and Armony (2017) and references therein. As more data from outpatient clinics become available, researchers can follow our approach to understand and improve the systems. For outpatient clinic managers, we provide a guideline on what data components they need to collect and how such data can be used to better understand and improve their systems. Our novel modeling approach can be easily generalized, which means it can be applied to other clinics in other countries and to other service systems with arrivals scheduled by appointments.

Second, we provide insights on what assumptions may be realistic when modeling the patient arrival process for outpatient clinics and how to check them using data. Understanding and incorporating human behavior into modeling outpatient clinics is becoming increasingly important. For example, Liu et al. (2010) and Liu and Ziya (2014) study scheduling decisions under patient no-shows and cancellations. In the clinic we analyze, we find that patients tend to be late more in the morning than in the afternoon. We also find that new patients are less likely to be no-shows but more likely to be late than repeat patients are. We believe our findings can motivate researchers and outpatient clinic managers to look for similar behavior in their systems and to develop models that incorporate such behavior.

Finally, we find that the main deviation from a regular deterministic arrival pattern (often assumed for appointment-generated arrival processes) is variability in the schedule itself. At first glance, viewing the schedule as random might appear inappropriate because, unlike call centers, where arrivals are generated exogenously, an appointment-generated schedule is endogenous, meaning that it is at least partly controlled by management. However, filling the master schedule is rarely straightforward (Liu et al. 2016), suggesting that it may be natural to view the final schedule as random, corresponding closely to random demand. An important managerial insight is that the schedule itself may be random and that it may be necessary to carefully model, monitor, and manage the schedule. It is evident that the master schedule is important, but it may not be evident that examining the schedules resulting from the master schedule as well as adherence to that schedule can also be important. We also observe that it is appropriate to regard the schedule as a stochastic process, evolving over time. To the best of our knowledge, this is the first study of an outpatient clinic to suggest that the schedule itself should be regarded as random and to characterize its stochastic structure.

2. Defining and Modeling the Daily Schedule

We now examine the schedule and arrival data for one clinic doctor over his 22 A.M. shifts. The arrivals planned for each day are given in a daily schedule, which has a specified number of arrivals in each of several evenly spaced 10-minute time intervals. Our schedule data are the 22 observed schedules for the doctor during his A.M. shifts. Even though much can be learned from consulting the appointment manager, we try to see what can be learned directly from the data. While conducting the data analysis, we confirmed our observations with clinic doctors and administrators.

2.1. The Evolution of a Schedule

The actual schedule for a given day evolves over time, typically starting many weeks before the specified day. We do not consider the scheduling process; instead, we consider the evolution of the resulting schedule. We regard the evolution of the schedule as a stochastic process, with additions and cancellations occurring randomly over time. For each day, we define the final schedule as its value at the end of the previous day.

In the left-hand panel in Figure 2, we illustrate the evolution of the daily cumulative number of patients scheduled over the previous year for the 22 days in the data set for 2013. The panel shows the specific appointment days as well, which are spread out between July and October.

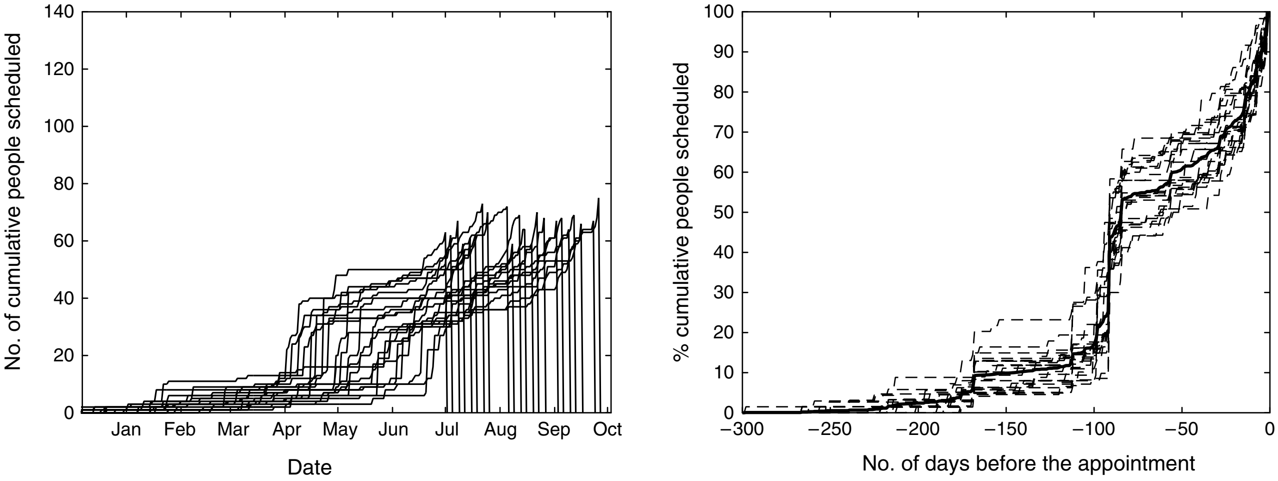
The right-hand panel in Figure 2 presents a useful alternative view, showing the percentage of the final schedule reached k days before the appointment data as a function of k. For all 22 days, 100% of the schedule is filled at $k = 0$. We see much less variability in the right-hand panel than in the left-hand panel. The percentage of the schedule reached 30 days before appears at $k = -30$. Especially revealing is the average of the 22 sets of percentage data, which is shown by the single thick line. From this average plot, we see jumps at regular intervals, especially around 90 days (3 months) before the appointment date. The right-hand plot in Figure 2 shows that about 24% of all appointments are made more than 93 days in advance, while about 30% are made between 93 and 84 days in advance (about 3 months). About 30% are made in the last month, while about 13% are made in the last week.

2.2. New and Repeat Visits

There is increasing interest in the delays from request to appointment, including how to determine panel sizes (the pools of potential patients) for doctors; see Green et al. (2007), Liu et al. (2010), Liu and Ziya (2014), and Zacharias and Armony (2017) and references therein. Unfortunately, our data set does not include a measure of the urgency or time sensitivity of each appointment, so we cannot determine whether patients are unable to get urgent appointments quickly enough. Fortunately, the data set does specify whether each scheduled arrival is a repeat visit or a new visit. Since 78% of all appointments are repeat visits, we conclude that the long intervals between the scheduling date and the appointment date do not imply that patients are failing to get urgent needs addressed promptly.

Figure 3 separately displays the evolution of the schedules for new and repeat visits, expanding upon the view in Figure 2. The figure panels show that this classification is important. Figure 3 specifically shows that only about 65% of new patients wait for more than a week for an appointment. The median number of days between the appointment scheduling date and the actual appointment date is 88 for repeat visits and 14 for new patients.

Figure 2. The Evolution of the Daily Cumulative Number of Patients Scheduled



Notes. *Left panel*: Evolution of the daily cumulative number of patients scheduled over the previous year for the 22 appointment days. The plot shows the specific appointment days as well, which are spread out between July and October. *Right panel*: The percentage of patients who are scheduled k days in advance for each of the 22 appointment days. The thick line indicates the average over the 22 appointment days.

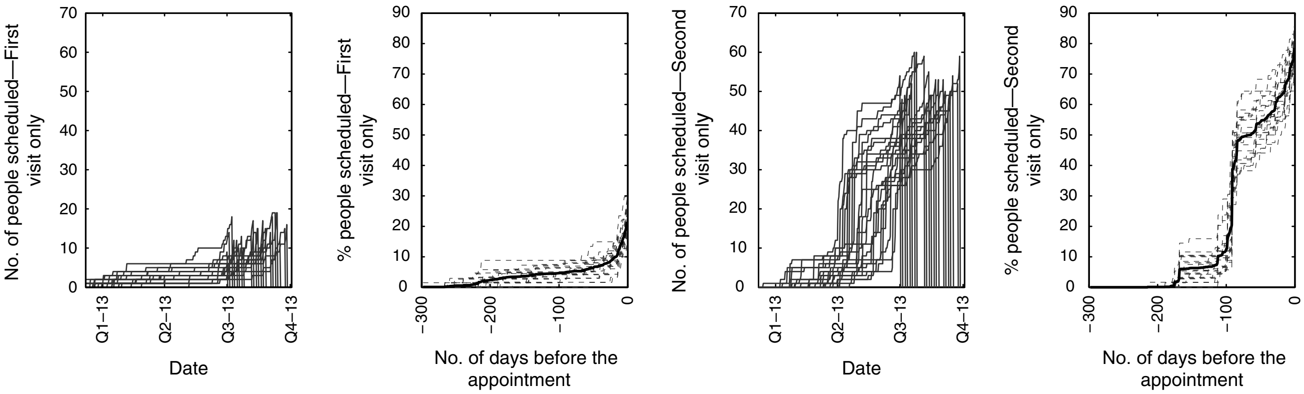
2.3. Inferring the Master Schedule

Recall that the master schedule is the framework designed for the schedule; it also called an appointment template. It usually specifies the total number of appointment slots for each day, the length of each appointment slot, and the number of patients to be scheduled for each slot (Liu et al. 2016). From the perspective of the eventual arrival process over each day, the evolution of the schedule should not matter much if the final schedule reaches the master schedule (or a schedule that is nearly deterministic and hence varies little from day to day). However, for the clinic, there is considerable variability in the realized schedules, so the evolution may matter.

We first define the schedule as the daily total plus the actual scheduled arrival times of all these patients. In particular, we define the schedule as its value at the end of the previous day, and we define arrivals on the same day as unscheduled arrivals. Given that definition, we next look for the underlying master schedule. The starting point for our data analysis is the 22 observed daily schedules. These are displayed in Table 1. Table 1 shows the number of patients scheduled for different 10-minute time slots (displayed vertically) over the A.M. shifts of 22 days (displayed horizontally). Each 10-minute time slot is specified by its start time.

Most appointment schedules today are designed and managed to fit into a master schedule, usually using a computerized appointment management system. However, it seems prudent to look at the actual schedules and infer the realized framework from the data. Not all variability occurs because of nonadherence to the schedule; rather, the schedules show that there is substantial variability in the schedule itself.

Figure 3. Evolution of the Daily Number of Patients Scheduled and the Percentage of Patients Scheduled k Days in Advance for Each of the 22 Appointment Days for New Patients (Left Two Panels) and Repeat Visits (Right Two Panels)



Note. The thick line indicates the average over the 22 appointment days.

Table 1. Number of Patients Scheduled in Each 10-Minute Time Slot (Displayed Vertically) During 22 Morning Shifts (Displayed Horizontally)

Time slot	22 days in July–October 2013																						Avg	Var	Var/Avg
7:50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00		
8:00	0	0	0	0	0	0	1	1	1	0	1	0	0	0	1	0	0	0	0	0	1	1	0.32	0.23	0.71
8:10	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.05	0.05	1.00
8:20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00		
8:30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00		
8:40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00		
8:50	3	4	5	4	4	4	4	4	4	1	3	2	1	4	2	4	4	2	4	5	4	3	3.41	1.30	0.38
9:00	3	4	2	3	3	2	3	3	3	3	3	2	2	2	2	3	4	3	2	3	4	2	2.77	0.47	0.17
9:10	3	3	3	2	2	2	4	2	2	3	2	3	2	3	3	3	2	2	3	2	3	3	2.59	0.35	0.13
9:20	2	2	4	2	3	2	3	2	2	3	3	3	2	3	2	3	3	3	3	2	3	2	2.59	0.35	0.13
9:30	3	2	3	4	3	3	4	3	3	3	3	3	1	3	2	2	2	2	3	3	3	3	2.77	0.47	0.17
9:40	3	3	3	2	2	2	2	3	3	2	2	3	2	3	2	2	2	2	3	2	2	2	2.36	0.24	0.10
9:50	3	3	3	3	2	3	3	3	3	3	3	2	2	3	3	3	3	3	2	2	3	3	2.77	0.18	0.07
10:00	3	2	3	3	2	3	2	3	2	3	3	3	3	3	3	3	4	4	3	3	3	3	2.91	0.28	0.10
10:10	3	3	3	3	3	3	3	3	2	2	3	3	3	3	3	3	3	3	3	3	3	3	2.91	0.09	0.03
10:20	2	3	3	3	3	3	3	2	3	3	2	3	3	3	3	2	3	2	3	4	3	3	2.82	0.25	0.09
10:30	3	2	3	3	3	2	4	2	3	2	3	3	3	3	3	2	3	3	2	4	3	3	2.82	0.35	0.12
10:40	3	1	3	3	3	1	3	2	3	2	3	3	2	3	2	1	3	2	3	3	3	2	2.45	0.55	0.22
10:50	2	3	3	3	1	2	3	2	3	3	3	2	3	3	3	3	3	3	2	3	3	3	2.68	0.32	0.12
11:00	3	2	3	2	3	2	3	2	2	4	4	4	2	3	3	3	3	3	3	4	3	4	2.95	0.52	0.18
11:10	3	3	3	1	3	3	3	3	2	3	3	2	3	2	1	3	2	3	3	3	3	3	2.64	0.43	0.16
11:20	2	3	3	3	3	3	3	3	3	3	3	3	3	2	2	3	3	3	3	3	3	4	2.91	0.18	0.06
11:30	3	3	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2	2	2	3	2	2.77	0.18	0.07
11:40	3	2	3	3	2	3	3	3	3	1	2	3	3	2	3	3	3	3	3	3	2	3	2.68	0.32	0.12
11:50	3	3	3	3	3	2	2	3	3	2	3	2	4	3	3	3	2	2	3	3	1	3	2.68	0.42	0.16
12:00	2	3	3	2	3	3	4	3	3	2	3	3	3	3	3	3	3	3	2	2	3	4	2.86	0.31	0.11
12:10	3	3	3	2	3	3	2	3	2	3	3	2	3	3	4	3	1	2	3	2	3	3	2.68	0.42	0.16
12:20	2	4	3	2	3	3	3	3	4	3	3	3	3	2	2	3	1	3	1	4	3	3	2.77	0.66	0.24
12:30	2	1	0	0	0	3	3	3	3	2	2	2	2	3	3	3	2	4	3	1	2	3	2.14	1.27	0.59
12:40	0	0	0	0	0	2	2	4	3	0	3	2	1	2	3	3	4	2	3	0	0	3	1.68	2.13	1.27
12:50	0	0	0	0	0	0	0	1	4	0	0	0	0	3	4	0	2	0	4	0	0	4	1.00	2.67	2.67
13:00	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.09	0.09	0.95
Daily total	63	62	67	59	61	62	73	70	72	59	69	64	59	70	68	67	68	64	69	66	67	75	66.09	21.32	0.32
[8:50, 12:20] total	60	61	67	59	60	57	67	60	61	57	63	60	56	62	57	61	60	58	59	65	64	64	60.82	9.77	0.16
All slot avg	2.0	2.0	2.2	1.9	2.0	2.0	2.4	2.3	2.3	1.9	2.2	2.1	1.9	2.3	2.2	2.2	2.2	2.1	2.2	2.1	2.2	2.4	2.07	1.73	0.84
All slot var	1.5	1.9	2.2	1.9	1.8	1.5	1.8	1.3	1.5	1.7	1.5	1.5	1.6	1.5	1.3	1.7	1.8	1.6	1.6	2.2	1.8	1.6	(across all days)		
All slot var/avg	0.7	1.0	1.0	1.0	0.9	0.8	0.8	0.6	0.6	0.9	0.7	0.7	0.8	0.6	0.6	0.8	0.8	0.8	0.7	1.1	0.8	0.7			
[8:50, 12:20] avg	2.7	2.8	3.0	2.7	2.7	2.6	3.0	2.7	2.8	2.6	2.9	2.7	2.5	2.8	2.6	2.8	2.7	2.6	2.7	3.0	2.9	2.9	2.76	0.42	0.15
[8:50, 12:20] var	0.2	0.6	0.3	0.5	0.4	0.4	0.4	0.3	0.4	0.5	0.2	0.3	0.5	0.3	0.4	0.4	0.7	0.3	0.4	0.7	0.4	0.4	(across all days)		
[8:50, 12:20] var/avg	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.1	0.1	0.2	0.1	0.1	0.2	0.1	0.2	0.1	0.3	0.1	0.2	0.2	0.1	0.1			

We next define what we mean by a master schedule. A general master schedule has batches of size β_j customers arriving at intervals τ_j after an initial time 0 for $1 \le j \le \nu$. Thus, the associated arrival times are

$$\psi_j \equiv \sum_{i=1}^{j-1} \tau_i \quad \text{for } 1 \le j \le \nu \text{ and } \psi_1 \equiv 0. \tag{1}$$

The framework has a total targeted number N_F and time T_F defined by

$$N_F = \sum_{j=1}^{\nu} \beta_j \quad \text{and} \quad T_F = \sum_{j=1}^{\nu-1} \tau_j = \psi_{\nu-1}. \tag{2}$$

A principal case is the *stationary framework*, with $\beta_j = \beta$ and $\tau_j = \tau$ for all j, which makes $N_F = \beta\nu$ and $T_F = (\nu - 1)\tau$, leaving the target parameter triple (β, τ, ν), but there often are variations in practice. In the more general model, it is important to consider alternative nonstationary schedules that might be used or contemplated to improve various measures of performance.

From Table 1, we infer that the master schedule is valid with $\tau = 10$ minutes. However, the scheduled arrivals in each time slot are not constant over different days or over different times on each day. Some shifts start as early as 8:00, and some end as late as 13:00. We observe that between 8:50 and 12:20, the average and average/variance of scheduled arrivals in each time slot across different days are comparable. Hence, for the A.M. shifts of the doctor in the endocrinology clinic, the stationary framework is roughly valid as an idealized model, with $\beta = 3$, $\tau = 10$ minutes and $\nu = 22$ and starting at 8:50 and ending at 12:20 (including the intervals [8:50, 9:00) and [12:20, 12:30), closed on the left and open on the right), which we refer to as the interval [8:50, 12:20]. The daily total for the stationary framework is $22 \times 3 = 66$, which matches the average daily total for the 22 days, even though the schedule is otherwise more variable.

On closer examination, we can see consistent structure in the schedule variability. First, we see that some days have higher daily totals, evidently because an effort is being made to respond to high demand. Second, we see random batch sizes in the slots over the entire shift. We discuss each of these features in turn.

2.4. Low- and High-Demand Service Systems

In general, it seems useful to classify service systems with arrivals by appointment into two categories. First, there are the low-demand service systems, for which it is challenging to fill a target schedule. For such service systems, the randomness in the schedule is a result of the random level of demand. We then might focus on the extent to which demand is adequate to fill the master schedule.

Second, there are the high-demand service systems, for which there is almost always ample demand, and often excess demand. In this case, the system may or may not actually respond to the excess demand, i.e., it may or may not schedule more than the normal workload to meet that excess demand. Of course, there can be more complicated scenarios in which a service system oscillates between the low-demand and high-demand modes.

If we identify the master schedule for the A.M. shifts as the 22 10-minute time slots in the interval [8:50, 12:20] in Table 1, then we observe that the daily totals within this interval are remarkably stable, having mean 60.82 and variance 9.77. In contrast, the full daily totals for the entire A.M. shifts are much more highly variable, having a variance of 21.19. From this observation, we infer that the doctor operates as a high-demand service system and that indeed he responds to excess demand on some but not all days. This conclusion is further confirmed by the observation that the extra patients tend to be scheduled outside (after) the main interval [8:50, 12:20]. Table 2 shows the distribution of the number of scheduled patients in these outside intervals N_o.

As further confirmation of the idea that overload appears outside the main time interval, we also see higher numbers in the first shift, at 8:50 (the interval [8:50, 9:00)); this suggests that at least some of the patients scheduled in the first interval, at 8:50, are scheduled in response to pressure to provide service to more patients than the usual number. We note that the first interval might be regarded as an overload period as well, though we choose not to do so. Moreover, the data show that the appointments in the outside interval (at the beginning and the end) were consistently made far closer to the actual appointment date than the other appointments were, with the median number of days between the appointment scheduling date and the actual appointment date 85 for the main interval and 14 for the outside interval.

In the online supplement, we elaborate on the scheduled appointments outside the main interval; we discuss how to measure the amount of excess demand and provide comparisons with three other shifts and discuss the impact of observed differences on the stochastic arrival process model we develop in this study.

2.5. Random Batch Sizes

Table 1 indicates that the number of patients scheduled for each 10-minute time slot is variable. Table 2 shows the distribution of the schedule within each time slot within the main interval. From Table 2, we conclude that it is reasonable to assume that the batch sizes in each of the time slots of the main time interval can be regarded as realizations of a random variable B_s, assuming values in the set $\{1, 2, 3, 4\}$ for any j. (We omit the value 5 because the frequency is so low, and we could also possibly omit the value 1 for the same reason.) In particular, we estimate the distribution as

$$P(B_s = k) = 0.03, 0.26, 0.63, 0.08, \quad 1 \le k \le 4, \tag{3}$$

so that

$$\begin{aligned} E[B_s] = 2.76, \quad E[B_s^2] = 8.02, \\ \mathrm{Var}(B_s) = 0.402, \quad \text{and} \quad SD(B_s) = 0.634, \end{aligned} \tag{4}$$

for all j. The variance is considerably less than the mean, so we can conclude that the distribution of B_s is much less variable than Poisson. The squared coefficient of variation (scv, or the variance divided by the square of the mean) is low as well, being $c_B^2 = 0.053$.

Table 2. Estimated Distribution of the Batch Sizes (B_s) Within the Main Interval [8:50, 12:20] and the Estimated Distribution of the Total Number of Scheduled Arrivals After the Main Interval (N_o)

	$\hat{P}(B_s = k)$					$\hat{P}(N_o = k)$										
Number k	1	2	3	4	5	0	1	2	3	4	5	6	7	8	9	10
Estimated distribution	0.03	0.26	0.63	0.08	0.00	0.14	0.09	0.14	0.05	0.05	0.14	0.09	0.00	0.14	0.00	0.18

2.6. Independence or Dependence Among Batch Sizes

In Section 2.5, we focused on the distribution of the batch size of the scheduled arrivals in any time slot within the main time interval on any day. We now consider the joint distribution of the batch sizes over successive time slots on the same day.

Let $B_{s,j}$ be the scheduled batch size in slot j, $1 \leq j \leq 22$, on a given day. For simplicity from a stochastic modeling perspective, it is natural to assume that the batch variables $B_{s,j}$ in successive slots j are independent, which corresponds to appointments being made independently for specific slots. However, it may be more realistic to assume that the appointments are primarily made with a specific day in mind and that the actual appointments are distributed approximately evenly over the day, with the person or system creating the schedule only partly in response to patient requests regarding specific time slots. Alternatively, appointments may overflow into nearby slots, which should also create positive correlation. Therefore, in any context, it is interesting to explore the dependence among the scheduled batch sizes $B_{s,j}$ for each day.

To illustrate, let N_S be the daily total of the schedule (focusing on the main interval [8:50, 12:20] with $\nu = 22$ slots); consider the case in which the distribution of B_s is independent of j. If the batch sizes are mutually independent, then

$$\mathrm{Var}(N_S) = \nu\,\mathrm{Var}(B_s). \tag{5}$$

In contrast, if we assume that the daily total is random and if we distribute it evenly among the slots, then we might have

$$B_s \approx \frac{N_S}{\nu} \quad \text{so that } \mathrm{Var}(N_S) = \nu^2\,\mathrm{Var}(B_s). \tag{6}$$

More generally, the dependence among the batch sizes might be usefully summarized by the correlations

$$\rho_{j_1,j_2} \equiv \mathrm{corr}(B_{s,j_1}, B_{s,j_2}) = \frac{\mathrm{cov}(B_{s,j_1}, B_{s,j_2})}{\sqrt{\mathrm{Var}(B_{s,j_1})\,\mathrm{Var}(B_{s,j_2})}}. \tag{7}$$

We propose a model that enables us to incorporate a range of possibilities in a parsimonious manner. We assume that

$$\rho_{j_1,j_2} = \rho_S, \quad -1 \leq \rho_S \leq 1, \quad \text{for all } j_1 \neq j_2. \tag{8}$$

We can then estimate the single pairwise correlation parameter ρ_S empirically in any given appointment setting.

Under assumption (8), we have

$$\begin{aligned}\sigma_S^2 \equiv \mathrm{Var}(N_S) &= \sum_{j=1}^{\nu}\sum_{k=1}^{\nu}\mathrm{Cov}(B_{s,j}, B_{s,k})\\ &= \nu\ \mathrm{Var}(B_s)[1 + (\nu-1)\rho_S].\end{aligned} \tag{9}$$

We thus estimate the correlation ρ_S in (8) by

$$\rho_S \equiv \frac{\mathrm{Var}(N_S) - \nu\,\mathrm{Var}(B_s)}{\nu\,\mathrm{Var}(B_s)(\nu-1)}, \tag{10}$$

where we use our estimates of $\mathrm{Var}(N_S)$ and $\mathrm{Var}(B_s)$. From Table 1, our estimate of $\mathrm{Var}(N_S)$ is 9.77; from (4), our estimate of $\nu\,\mathrm{Var}(B_s)$ is $22 \times 0.402 = 8.80$. We thus estimate that ρ_S is $0.97/185 = 0.0052$, which is sufficiently small that we consider the i.i.d. model reasonable.

2.7. Outside the Main Time Interval

It remains to specify arrivals scheduled outside the main time interval. Since the average total outside is only about 10% of the full daily total and since we do not have a great amount of data overall, we will not try to develop a high-fidelity model. Based on the limited data provided by Tables 1 and 2, we allocate the total number of scheduled arrivals outside (after) the main interval according to the distributions specified in Table 2. If the total number to be scheduled outside the main interval is seven or fewer, then we divide the number into two parts, putting the larger or equal number in the first slot and the smaller or equal number in the second slot. If the total number is eight or more, we divide the total into three parts, as evenly as possible, and put the numbers in decreasing order in the first three slots after the main interval.

2.8. Summary of the Schedule Model

In summary, the clinic data clearly indicate a well-defined structured framework, provided that we focus on a main time interval [8:50, 12:20] containing 22 slots. The scheduled numbers in these slots can be regarded as i.i.d. random variables distributed as the random variable B_s, as in (3). Our analysis in Section 2.6 supports regarding these slot numbers as mutually independent.

Our doctor evidently experiences high demand. As stipulated in Section 2.7, we allocate the totals randomly according to the distributions in Table 2, and we distribute them in a balanced, decreasing order over the outside intervals. Since the numbers outside are smaller, we devote less effort to developing a high-fidelity model for that part.

Only about 10% of the mean of the daily totals (66) is due to the arrivals scheduled outside the main interval (the mean inside is 60.8), while the variance 21.2 in the daily totals is primarily a result of the random occurrence of arrivals scheduled outside the main interval because the variance inside is 9.77. (See Equation (16) for a more precise statement.) Thus, we tentatively conclude that the greatest contributor to the overall variability of the schedules for the doctor in our study is the inconsistent response to extra demand. By examining both the scheduled and the realized arrivals for the other 15 doctors in the clinic, we find that this observation applies to all the other doctors as well: see the online supplement and Figures 1–3 and Figures 4–11 in our longer, more detailed study (Kim et al. 2015a). To draw a firm conclusion, we would need to consider data on the original demand, i.e., requests for appointments, including ones that were not satisfied or that were moved to another day.

3. Adherence to the Schedule: Patient No-Shows and Unscheduled Arrivals

We now come to the question of adherence to the schedule. The level of adherence converts the schedule into the actual arrival process. We identify three familiar forms of additional randomness in the model: (i) no-shows, (ii) extra unscheduled arrivals, and (iii) lateness or earliness. We first focus on the no-shows and the unscheduled arrivals, which together determine how the scheduled daily number of arrivals is translated into the actual daily total number of arrivals. In Section 4, we focus on lateness or earliness, which each have a significant impact on the pattern of actual arrivals over the day.

3.1. No-Shows

The no-shows are the scheduled arrivals who do not actually arrive. Instead of the number of actual arrivals in time slot j on a given day, which we denote by $B_{a,j}$, we begin by focusing on the number from among the $B_{s,j}$ arrivals who were scheduled to arrive in slot j on that day who did arrive *at some time on that appointment day*, which we denote by $B_{a\,|\,s,j}$, which necessarily satisfies the inequalities

$$0 \le B_{a\,|\,s,j} \le B_{s,j}, \quad \text{for all } j. \tag{11}$$

The no-shows in slot j are thus defined as

$$B_{n,j} \equiv B_{s,j} - B_{a\,|\,s,j}. \tag{12}$$

These are shown in Table 3.

Table 3 shows that no-shows are rarer than in many other appointment systems: the number of no-shows ranges from 2 to 10 per day, with an average of 5.45 per day. The overall proportion of no-shows is 5.45/66.09, or 8.2%.

In general, we might try to model the no-shows carefully, as we did the schedule batch sizes $B_{s,j}$, but here, we simply assume that each scheduled patient fails to arrive in each slot on each day with probability $\delta = 0.082$, independently of all other patients. Overall, in the model, the total number of no-shows would have a binomial distribution with parameters equal to the total number, say n, of scheduled patients over all days and with probability $p = \delta = 0.082$, which would make the distribution approximately Poisson, with variance slightly less than the mean. Table 3 shows that the observed sample variance of the average number of no-shows is 6.35, which is only slightly greater than the overall average of 5.45. Hence, we conclude that the model with i.i.d. Bernoulli no-shows is well supported by the data.

3.2. Unscheduled Arrivals

Some medical services have significant proportions of both unscheduled and scheduled arrivals. However, there are relatively few unscheduled arrivals at the clinic we study. As indicated before, we define them as arrivals that are scheduled on the same day (after the end of the previous day). On average, there are 2.18 unscheduled patients per day, among whom 1.95 arrived. In the online supplement, Table 11 shows all additional unscheduled arrivals, and Table 12 shows the additional unscheduled arrivals that actually arrived. The total number of unscheduled arrivals (that arrived) on all 22 days is 43. Table 12 shows that the total daily number of unscheduled arrivals exceeds 3 on only two days, with values of 4 and 7. The one exceptional day is evidently responsible for the variance for all days, 2.43, being larger than the mean. The unscheduled arrivals are more likely to be outside the main time interval, which is consistent with our interpretation of outside the main time interval being a time for overload.

Paralleling our previous modeling, we could represent the daily total number of unscheduled arrivals within the main time interval as Poisson with mean 1.55 and those outside the main interval as Poisson with mean 0.40. We could then distribute those arrivals randomly (uniformly) within the respective time periods. With larger numbers, we might try more careful modeling. However, in general, some sort of Poisson process is natural for unscheduled arrivals because they are likely to be a result of individual people making independent decisions.

3.3. Daily Totals

We now examine the impact of no-shows and unscheduled arrivals on the actual daily totals of arrivals. Let N_A, N_S, N_N, and N_U be the random daily total numbers of actual arrivals, scheduled arrivals, no-shows, and unscheduled arrivals, respectively. In general, we have the basic flow conservation formula

$$N_A = N_S - N_N + N_U. \tag{13}$$

Combining the summary data from Tables 1, 3, and 12 (Table 12 is in the online supplement), we see that the means are

$$\begin{aligned} E[N_A] &= E[N_S] - E[N_N] + E[N_U] \\ &= 66.1 - 5.5 + 2.0 = 62.6. \end{aligned} \tag{14}$$

Table 3. Number of No-Shows ($B_{n,j} \equiv B_{s,j} - B_{a|s,j}$) for Each 10-Minute Time Slot j (Displayed Vertically) During 22 Morning Shifts (Displayed Horizontally)

Time slot	22 days in July–October 2013																						Avg	Var	Var/ Avg
7:50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00		
8:00	0	0	0	0	0	0	1	1	1	0	1	0	0	0	1	0	0	0	0	0	1	1	0.32	0.23	0.71
8:10	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.05	0.05	1.00
8:20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00		
8:30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00		
8:40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00		
8:50	1	0	0	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0.23	0.28	1.23
9:00	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0.18	0.16	0.86
9:10	0	0	0	1	0	0	2	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0.27	0.30	1.11
9:20	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0	0	1	0	0	0	0	0.18	0.25	1.38
9:30	0	0	0	2	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.18	0.25	1.38
9:40	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0.18	0.25	1.38
9:50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.00		
10:00	0	0	0	0	0	0	0	0	0	1	0	2	1	0	0	0	1	0	1	0	0	0	0.27	0.30	1.11
10:10	0	0	0	1	0	0	0	1	0	0	1	1	0	0	0	0	1	0	0	0	2	0	0.32	0.32	1.01
10:20	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0.23	0.18	0.81
10:30	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0.18	0.16	0.86
10:40	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	1	0.23	0.18	0.81
10:50	0	0	1	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	2	0.27	0.30	1.11
11:00	0	0	0	1	1	0	0	0	0	1	1	0	1	0	0	1	0	0	0	1	0	0	0.32	0.23	0.71
11:10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	1	0.18	0.16	0.86
11:20	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	1	0	0	1	0	0	0.23	0.18	0.81
11:30	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	1	1	0	0	0	0	0.23	0.18	0.81
11:40	0	0	1	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0.23	0.18	0.81
11:50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0.14	0.12	0.90
12:00	0	0	1	0	0	0	1	0	1	1	0	1	2	0	0	1	2	0	0	0	1	1	0.55	0.45	0.83
12:10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0.05	0.05	1.00
12:20	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0.09	0.09	0.95
12:30	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0.14	0.12	0.90
12:40	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0.09	0.09	0.95
12:50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0.05	0.05	1.00
13:00	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.09	0.09	0.95
Daily total	3	2	6	8	2	2	10	7	5	4	6	10	6	2	5	5	6	5	4	7	5	10	5.45	6.35	1.17
[8:50, 12:20] total	2	2	6	8	1	2	8	4	4	4	4	10	6	1	4	5	6	5	4	7	4	7	4.73	5.64	1.19
All slot avg	2.0	2.0	2.2	1.9	2.0	2.0	2.4	2.3	2.3	1.9	2.2	2.1	1.9	2.3	2.2	2.2	2.2	2.1	2.2	2.1	2.2	2.4	0.17	0.17	1.00
All slot var	1.5	1.9	2.2	1.9	1.8	1.5	1.8	1.3	1.5	1.7	1.5	1.5	1.6	1.5	1.3	1.7	1.8	1.6	1.6	2.2	1.8	1.6	(across all days)		
All slot var/avg	0.7	1.0	1.0	1.0	0.9	0.8	0.8	0.6	0.6	0.9	0.7	0.7	0.8	0.6	0.6	0.8	0.8	0.8	0.7	1.1	0.8	0.7			
[8:50, 12:20] avg	2.7	2.8	3.0	2.7	2.7	2.6	3.0	2.7	2.8	2.6	2.9	2.7	2.5	2.8	2.6	2.8	2.7	2.6	2.7	3.0	2.9	2.9	0.21	0.21	0.98
[8:50, 12:20] var	0.2	0.6	0.3	0.5	0.4	0.4	0.4	0.3	0.4	0.5	0.2	0.3	0.5	0.3	0.4	0.4	0.7	0.3	0.4	0.7	0.4	0.4	(across all days)		
[8:50, 12:20] var/avg	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.1	0.1	0.2	0.1	0.1	0.2	0.1	0.2	0.1	0.3	0.1	0.2	0.2	0.1	0.1			

We see that the final mean daily number of arrivals $E[N_A] = 62.6$ is only about 5% less than the mean scheduled daily number $E[N_S] = 66.1$. Hence, from the perspective of the daily totals, there is strong adherence to the schedule.

Moreover, we see that the variability of the daily number of arrivals N_A is primarily caused by the variability of the schedule. The sample variances of the four daily numbers were

$$\begin{aligned} \mathrm{Var}(N_A) &= 17.4, \quad \mathrm{Var}(N_S) = 21.3, \\ \mathrm{Var}(N_N) &= 6.4, \quad \text{and} \quad \mathrm{Var}(N_U) = 2.4. \end{aligned} \tag{15}$$

Note that the estimated variances are ordered by $\mathrm{Var}(N_A) < \mathrm{Var}(N_S)$. The dispersions (sample variance divided by the sample mean) are similarly ordered as follows:

$$\begin{aligned} \mathrm{Var}(N_A)/E[N_A] &= 17.4/62.6 = 0.278 \\ &< 0.322 = 21.3/66.1 = \mathrm{Var}(N_S)/E[N_S]. \end{aligned} \tag{16}$$

4. The Arrival Pattern Over the Day: Patient Nonpunctuality

We now shift our attention to the pattern of arrivals over each day, given the daily totals. Here, "pattern" primarily means whether each patient arrives before or after the appointment time (earliness or lateness), but it might also mean systematic time dependence of the schedule, the no-shows, or the unscheduled arrivals over the day.

Table 4. Average Numbers of Scheduled Arrivals for Each 30-Minute Interval Within the Main 3.5-Hour Time Interval as Well as the Proportions of No-Shows and Lateness and the Average Earliness (X^-), Lateness (X^+), and Overall Deviation (X), Plus 95% Confidence Intervals

Interval	Avg no. scheduled	% no-show	% late	% (late > 15 min)	Avg(X^+)	Avg(X^-)	Avg(X)
[8:50, 9:20)	8.8 ± 0.7	7.9 ± 4.8	21.2 ± 6.9	12.3 ± 5.5	35.8 ± 18.7	−25.8 ± 2.7	−11.4 ± 6.9
[9:20, 9:50)	7.7 ± 0.5	6.9 ± 4.6	16.7 ± 6.1	4.8 ± 3.4	24.1 ± 25.4	−35.7 ± 5.6	−25.7 ± 5.2
[9:50, 10:20)	8.6 ± 0.4	6.8 ± 4.4	15.0 ± 6.7	6.4 ± 3.4	20.3 ± 10.9	−38.8 ± 5.2	−30.2 ± 6.5
[10:20, 10:50)	8.1 ± 0.6	7.9 ± 3.2	17.6 ± 5.0	3.3 ± 2.9	9.7 ± 4.9	−45.0 ± 7.3	−34.5 ± 6.0
[10:50, 11:20)	8.3 ± 0.5	9.0 ± 3.9	13.6 ± 4.4	5.4 ± 3.9	18.4 ± 11.2	−48.6 ± 9.1	−39.2 ± 8.7
[11:20, 11:50)	8.4 ± 0.3	7.9 ± 3.7	10.4 ± 4.7	3.9 ± 3.0	16.0 ± 6.4	−61.2 ± 9.1	−53.3 ± 9.4
[11:50, 12:20)	8.2 ± 0.5	9.3 ± 4.1	9.5 ± 5.4	3.8 ± 3.5	12.7 ± 6.6	−58.2 ± 9.5	−51.7 ± 9.8
[8:50, 12:20)	58.0±1.3	8.0±1.7	15.0±1.5	5.8±1.6	21.3±5.6	−44.9±3.0	−34.9±2.9

4.1. The Big Picture of the Daily Pattern

Table 4 provides the details of the big picture for the time interval [8:50, 12:20]. The first four columns of Table 4 show the average numbers scheduled, the percentage of no-shows, the percentage late, and the percentage late by more than 15 minutes by half-hour intervals over the A.M. shift, while the first four columns of Table 5 separately show the same summary statistics for new and repeat patients; these statistics are significantly different. Table 4 shows that the scheduled numbers and the no-shows are remarkably stable over time. As we have observed in previous sections, the main irregularity in the schedule occurs as a result of occasional overload scheduled outside these time intervals.

However, we see a different pattern in the lateness or earliness, as shown in the last four columns of Table 4. Specifically, Table 4 shows the percentage of patients arriving late, the average of the lateness X^+ among those patients arriving late, the average of the earliness X^- among those patients arriving early, and the overall average lateness X (these values are negative when the patient is early). Table 4 shows that the likelihood of lateness and the expected value of lateness tend to decrease over the day. In particular, we see that on average, 15% of the patients are late (arrive after the appointment time) each day, with an average lateness of $E[X^+] = 21$ minutes, but the percentage decreases over the day, from 21.2% in the first half hour to 9.5% in the last half hour. Meanwhile the average amount of lateness among these late patients, $E[X^+]$, decreases from 35.8 to 12.7 minutes. In general, Table 4 shows that patients tend to arrive early, rather than late. This again reflects strong adherence to the schedule.

4.2. Toward a Model of the Deviations

We now look closer into the deviations of the actual arrival times from the scheduled arrival times. Figure 4 shows the *empirical cumulative distribution functions* (ecdfs) for the lateness for each of the half-hour time slots in Table 4. Figure 4 shows that the lateness consistently decreases over the day in the sense that each successive ecdf is stochastically less than the one before; see Ross (1996, Section 9.1). (One ecdf is stochastically less than or equal to another if the entire ecdf lies *above* the other; e.g., the stochastically largest ecdf (with the most lateness) falls below all others and occurs in the first half hour.)

We now create a model of patient lateness (or earliness). The model has each scheduled arrival arrive at a random deviation from its scheduled arrival time. Let the arrivals scheduled to arrive at each time be labeled in some determined order, independent of the actual arrival time. We let the kth arrival among the scheduled arrivals in time slot j (at time ψ_j in (1)) actually occur at time

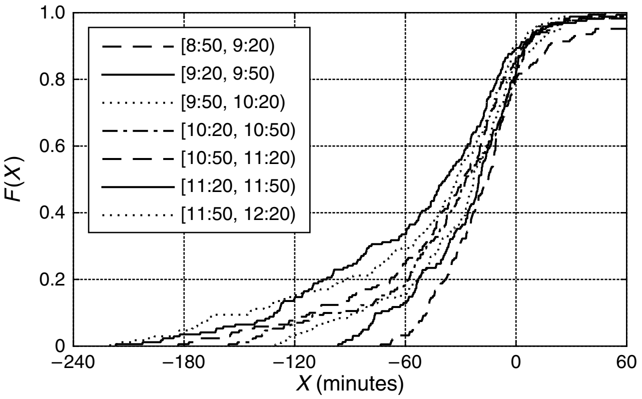
$$A_{j,k} = \psi_j + X_{j,k} = \sum_{i=1}^{j-1} \tau_i + X_{j,k}, \qquad (17)$$

where $X_{j,k}$ are mutually independent random variables, independent of the schedule (assuming arrivals

Table 5. Average Numbers for New and Repeat Patients for the Main Interval and Outside of the Interval as Well as the Proportions of No-Shows and Lateness and the Average Earliness (X^-), Lateness (X^+), and Overall Deviation (X), Plus 95% Confidence Intervals

Interval	Avg no. scheduled	% no-show	% late	% (late > 15 min)	Avg(X^+)	Avg(X^-)	Avg(X)
New	14.2 ± 1.3	5.5 ± 2.4	22.2 ± 4.4	7.7 ± 3.4	23.2 ± 9.7	−34.2 ± 4.4	−21.2 ± 3.9
New—[8:50, 12:30)	13.7 ± 1.3	5.7 ± 2.5	23.1 ± 4.6	8.0 ± 3.5	23.2 ± 9.7	−33.9 ± 4.7	−20.5 ± 4.1
New—outside	0.5 ± 0.3	0	0	0		−42.0 ± 27.9	−42.0 ± 27.9
Repeat	51.9 ± 2.1	8.8 ± 1.8	11.8 ± 1.7	4.7 ± 1.8	18.7 ± 5.8	−49.3 ± 3.3	−41.2 ± 3.6
Repeat—[8:50, 12:30)	47.1 ± 1.7	8.3 ± 1.9	12.1 ± 1.7	4.7 ± 1.8	18.8 ± 5.8	−48.7 ± 2.9	−40.4 ± 3.1
Repeat—outside	4.8 ± 1.5	16.9 ± 11.6	7.0 ± 6.8	4.0 ± 5.7	14.6 ± 12.9	−62.6 ± 24.1	−59.9 ± 25.3

Figure 4. Lateness Empirical Cumulative Distribution Functions in Each of the 30-Minute Intervals



are acting independently), and where $X_{j,k}$ is distributed as the random variable X_j with *cumulative distribution function* (cdf)

$$F_j(x) \equiv P(X_j \le x), \quad -\infty < x < +\infty. \tag{18}$$

We allow X_j to assume both positive and negative values, representing arriving late and arriving early, respectively.

The ecdfs in Figure 4 can be regarded as estimates of the cdf F_j, and we use the same cdf F_j for all three 10-minute time slots j in the specified half hour. For a simple model, we might want a single cdf F, but Table 4 and Figure 4 present strong evidence that F_j should be allowed to depend on j, at least to some extent.

Finally, we note that it may be useful to incorporate constraints on the arrival times at the beginning and the end of the time period. We might replace $A_{j,k}$ with the constrained version

$$A^c_{j,k} \equiv \max\{0, \min\{T_F, A_{j,k}\}\}. \tag{19}$$

To generate concrete stochastic models, we suggest fitting $P(X_j > 0)$ to the observed proportion of lateness in the half hour containing j and then fitting distributions to the observed values of lateness X^+ or earliness X^- separately. The lateness probability estimates are given directly in Table 4. If, instead, we use estimates of the cdf F of earliness or lateness, we would use the ecdfs, denoted by $\hat{F}(x)$, to generate the model cdfs of X^+ and X^-, letting

$$\begin{aligned} F_{X^+}(x) &\equiv P(X \le x \mid X > 0) \equiv \frac{\hat{F}_j(x) - \hat{F}_j(0)}{1 - \hat{F}_j(0)} \quad \text{and} \\ F_{X^-}(x) &\equiv P(X \le -x \mid X \le 0) \equiv \frac{\hat{F}_j(-x)}{\hat{F}_j(0)}, \quad x \ge 0. \end{aligned} \tag{20}$$

Figure 4 suggests it should be possible to use elementary parametric models. We show the results of fitting exponential distributions to X^+ and X^- over different hours in Figure 5. Figure 5 shows that the estimated scv c^2 is less than 1 for X^- and greater than 1 for X^+. Given the limited data, the exponential fit for X^- might be judged adequate, but we might also want to allow for greater variability in the lateness. We provide for that by considering a two-moment hyperexponential (mixture of two exponentials, with $c^2 > 1$ and balanced means, as in Whitt 1982) in Figure 6.

Given that we have specified the cdf's F_j, we have completed construction of a full stochastic model of the arrival process that can be used to simulate arrivals to the clinic.

4.3. Comparing Arrivals to the Schedule

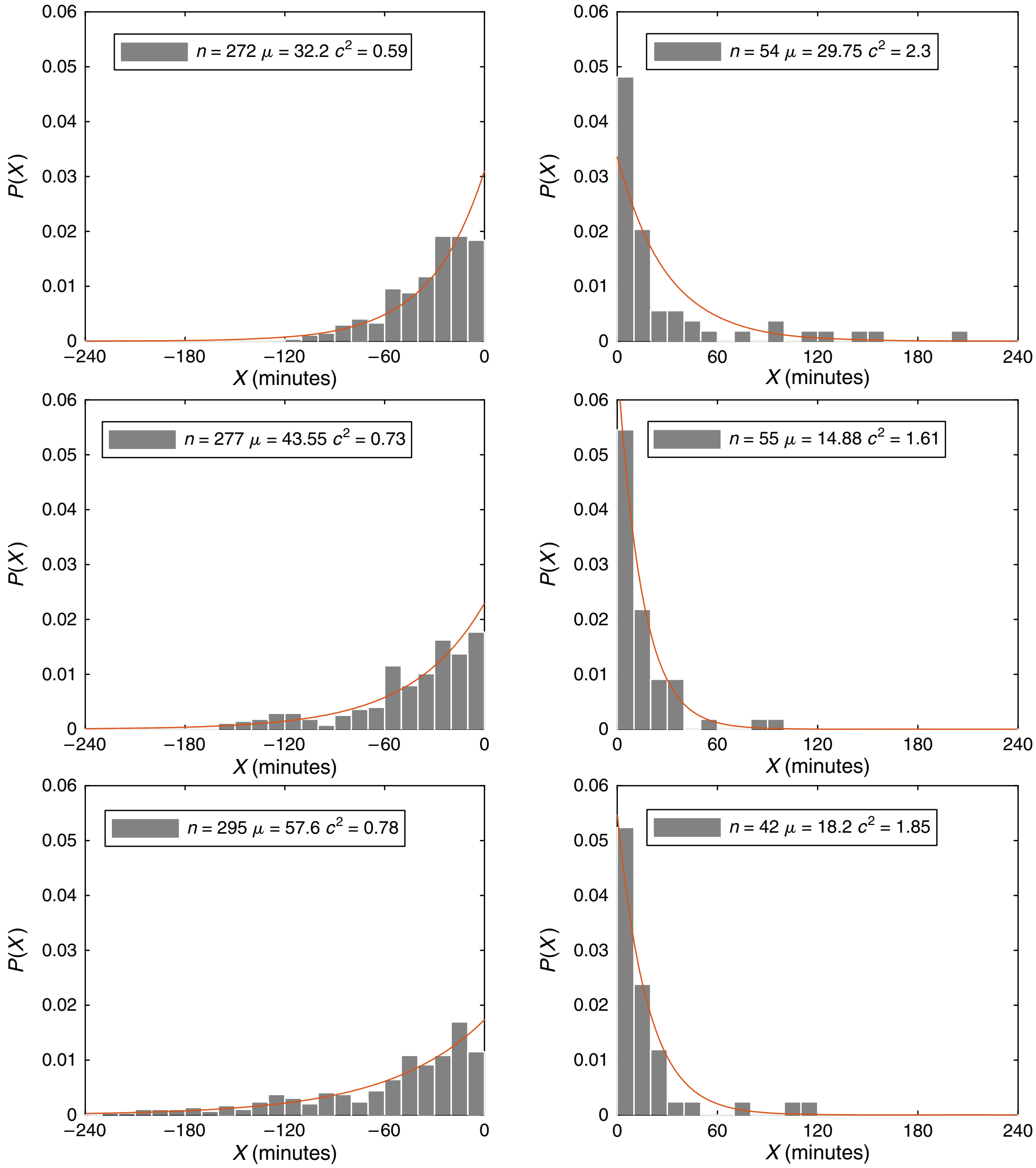
We now directly compare the realized arrivals to the schedule. Table 13 in the online supplement shows the difference between the numbers scheduled for the time slot and the numbers that arrived in that slot for each slot during the 22 days. The difference is often large, which we have seen must be primarily a result of deviations from the scheduled arrival times, especially earliness. Figures 7 and 8 provide summary views.

Let $S(t)$ and $A(t)$ count the number of scheduled and actual arrivals up to time t. Figure 7 shows the histograms of the 22 observed values of the counting processes $S(t)$ and $A(t)$ for a few values of t: 10 A.M., 11 A.M., 12 P.M., and 1 P.M. In particular, Figure 7 exposes systematic effects and shows the variability. Based on the figure, the cumulative number of arrivals scheduled is in general smaller than the cumulative number of actual arrivals for 10 A.M. and 11 A.M., but they are about the same at 12 P.M. and become smaller at 1 P.M. We have seen that this is caused by the earliness of patient arrivals and patient no-shows.

Figure 8 summarizes the data by plotting the average numbers of scheduled and actual arrivals for each of the 10-minute time slots within the 22 A.M. shifts. Figure 8 also shows linear rate functions fit by least squares to the 22 averages of the scheduled and actual arrivals for each of the 22 10-minute time slots within the main time interval (the solid lines). As should be expected, we see that the estimated rate function for the schedule within the main time interval is constant but that the estimated rate function of the actual arrivals is decreasing because of the tendency for patients to arrive early.

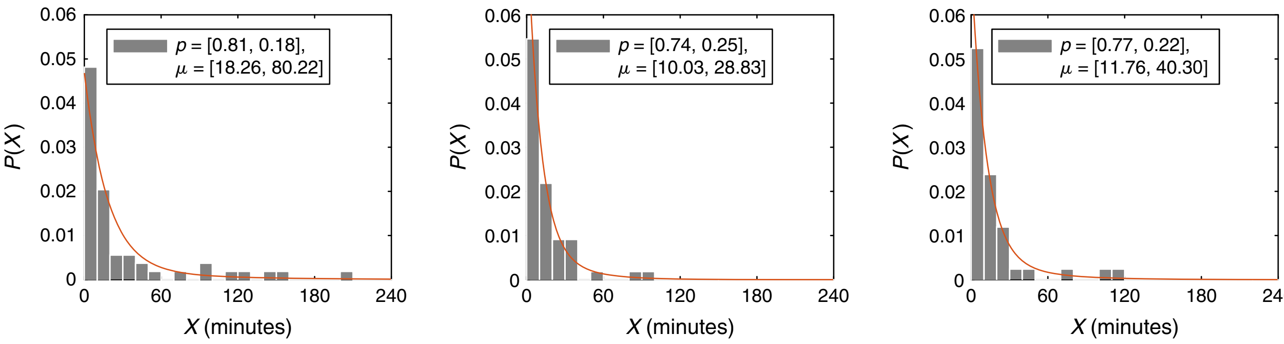
Finally, Figure 8 shows an additional continuous piecewise-linear estimated arrival rate function (the dotted lines) for the arrivals over the three intervals of the A.M. shift. This dotted line has an extra linear piece before the main interval to account for the earliness. We will use this construction as the arrival rate resulting from the schedule in the main interval in the simple model constructed in Section 5.3.

Figure 5. (Color online) Earliness (X^-) and Lateness (X^+) Histograms and Associated Exponential Fits



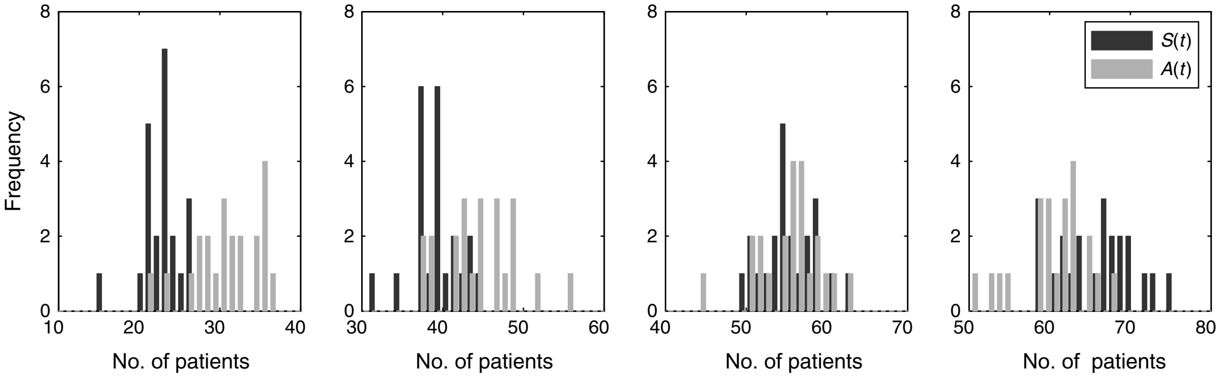
Note. Top to bottom: scheduled arrivals in [9, 10), [10, 11), and [11, 12).

Figure 6. (Color online) Lateness (X^+) Histograms and Associated Hyperexponential (H_2) Fits



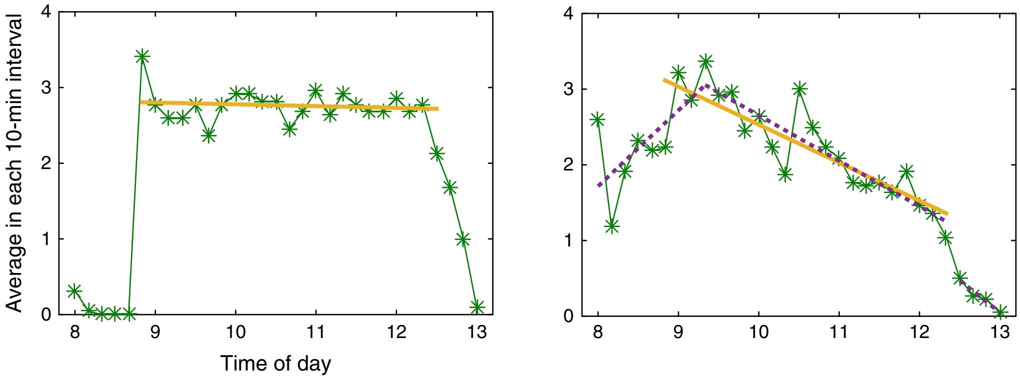
Note. Left to right: scheduled arrivals in [9, 10), [10, 11), and [11, 12).

Figure 7. Histograms of the Counting Processes $S(t)$ and $A(t)$ at Four Different Times



Notes. From left to right: 10 A.M., 11 A.M., 12 P.M., and 1 P.M.

Figure 8. (Color online) Plots of the Average Numbers of Scheduled (Left) and Actual (Right) Arrivals in Each of the 22 10-Minute Intervals in the Interval [8:50, 12:20] and Their Fitted Lines



5. Stochastic Models for the Arrival Process

In this section, we give concise mathematical representations of the stochastic counting processes $S(t)$ and $A(t)$, counting the number of scheduled and actual arrivals up to time t in the A.M. shift, defined in terms of the model elements developed in previous sections.

The number of scheduled arrivals up to time t can first be expressed as the sum

$$S(t) = \sum_{j=1}^{k} B_{s,j}, \quad \psi_k \le t < \psi_{k+1}, \, k \ge 0, \tag{21}$$

for all t, for ψ in (1) and the batch sizes $B_{s,j}$. According to the model in Section 2, $B_{s,j}$ should be i.i.d. random variables with distribution in (3) inside the main time interval and distributed outside according to Section 2.7.

Let $A_S(t)$ count the number of scheduled arrivals that actually arrive up to time t. To define $A_S(t)$, let the scheduled arrivals in each arrival epoch j (at time ψ_j) be ordered in some definite manner not having to do with their actual arrival time. Let $I_{j,k} = 1$ if scheduled arrival k at time ψ_j actually arrives on that day and let $X_{j,k}$ be the deviation of the actual arrival time from the scheduled time. If $X_{j,k} > 0$, the arrival is late; otherwise, the arrival is early. (For simplicity in labeling, we have variables $X_{j,k}$ even when $I_{j,k} = 0$, but they will play no role.) We combine these two random features with the indicator random variable $I_{j,k}(t)$, defined by

$$I_{j,k}(t) \equiv 1_{\{I_{j,k}=1,\, X_{j,k} \le t\}}, \quad -\infty < t < \infty,\ 1 \le k \le B_{s,j},\ j \ge 0. \tag{22}$$

Given these definitions, we can write

$$A_S(t) = \sum_{j=1}^{\infty} \sum_{k=1}^{B_{s,j}} 1_{\{I_{j,k}=1,\, X_{j,k} \le t - \psi_j\}} = \sum_{j=1}^{\infty} \sum_{k=1}^{B_{s,j}} I_{j,k}(t - \psi_j), \tag{23}$$

for $-\infty < t < +\infty$, where ψ_j is defined in (1). We may have $A_S(t) > 0$ for $t < 0$ because of early arrivals.

Let $A_U(t)$ ($A(t)$) count the number of unscheduled (all) arrivals by time t. Then we have

$$A(t) = A_S(t) + A_U(t), \quad \text{for all } t. \tag{24}$$

From Section 3.2, for the clinic, A_U would be independent of A_S, having two independent Poisson-based components, one for inside the main time interval and another for outside.

5.1. Conditional Means and Variances

Now suppose that the schedule is known; i.e., we know $B_{s,j}$ for all j, as would be the case at the end of the previous day in the clinic. Let the information about the schedule ($B_{s,j}$ for all j) be denoted by $\mathscr{S}$.

Since the ordering on k for each j is totally arbitrary, it is natural to assume that the joint distribution of $(I_{j,k}, X_{j,k})$ is independent of k for each j, and we make that assumption. The conditional cumulative arrival rate function for the scheduled arrivals given the schedule is then simply the conditional expected value

$$\Lambda_S(t \mid \mathscr{S}) \equiv E[A_S(t) \mid \mathscr{S}] = \sum_{j=1}^{\infty} B_{s,j} p_j(t), \qquad (25)$$

$-\infty < t < +\infty$, where

$$p_j(t) \equiv E[I_{j,k}(t-\psi_j)] = P(I_{j,k} = 1, X_{j,k} \le t - \psi_j) \\ = (1-\delta)F_j(t-\psi_j), \qquad (26)$$

with $F_j(t) \equiv P(X_{j,k} \le t)$, which is independent of k. As usual, the associated arrival rate function $\lambda_S(t \mid \mathscr{S})$ is the derivative with respect to t of the cumulative arrival rate function $\Lambda_S(t \mid \mathscr{S})$; i.e.,

$$\lambda_S(t \mid \mathscr{S}) = \sum_{j=1}^{\infty} B_{s,j}(1-\delta) f_j(t-\psi_j), \qquad (27)$$

where f_j is the probability density function (pdf) associated with the cdf F_j. The associated conditional variance is

$$V_S(t \mid \mathscr{S}) \equiv \mathrm{Var}(A_S(t) \mid \mathscr{S}) = \sum_{j=1}^{\infty} B_{s,j}^2 p_j(t)(1-p_j(t))$$

for $p_j(t)$ in (26).

5.2. The Total Mean and Variance of $A(t)$

The total arrival rate function is then

$$\begin{aligned}\Lambda(t) &\equiv E[A(t)] = E[A_S(t)] + E[A_U(t)] \\ &= E[\Lambda_S(t|\mathscr{S})] + E[A_U(t)] \\ &= \sum_{j=1}^{\infty} E[B_{s,j}](1-\delta)F_j(t-\psi_j) + E[A_U(t)]. \qquad (28)\end{aligned}$$

Applying the conditional variance formula, assuming that the random variables $B_{s,j}$ are mutually independent, the associated variance is

$$\begin{aligned}&\mathrm{Var}(A(t)) \\ &\quad = \mathrm{Var}(A_S(t)) + \mathrm{Var}(A_U(t)) \\ &\quad = \mathrm{Var}(E[A_S(t \mid \mathscr{S})]) + E[\mathrm{Var}(A_S(t) \mid \mathscr{S})] + \mathrm{Var}(A_U(t)) \\ &\quad = \sum_{j=1}^{\infty} \mathrm{Var}(B_{s,j})[(1-\delta)F_j(t-\psi_j)]^2 \\ &\qquad + \sum_{j=1}^{\infty} E[B_{s,j}^2]p_j(t)(1-p_j(t)) + \mathrm{Var}(A_U(t)). \qquad (29)\end{aligned}$$

5.3. A Parsimonious Simplified Arrival Process Model

We divide the overall time interval [8:00, 13:00] into two parts: before and after 12:30. We let D_F be the daily total during the final interval [12:30, 13:00]. For each day, we let D_F be distributed as in Table 2, which makes the overall mean number in [12:30, 13:00] 4.82. We then distribute the D_F arrivals among the intervals, as indicated in Section 2.7.

We let D_I be the random daily total for the initial interval [8:00, 12:20]. We let $E[D_I] = 66.1 - 4.8 = 61.3$, making it coincide with the observed average total of 66.1 in Table 1. We let the variance coincide roughly with the variance of the schedule inside the interval in Table 1, so that $\mathrm{Var}(D_I) = 10.0$. We can use a Gaussian distribution (rounded to the nearest integer) with this estimated mean and variance. Alternatively, we can fit a binomial distribution with parameter pair (n, p) to this mean and variance, yielding two equations with two unknowns: $E[D_I] = np = 61.3$ and $\mathrm{Var}(D_I) = np(1-p) = 10$, so that $(1-p) = 10/61.3 = 0.163$ and $n = 61.3/0.837 = 73.2$, rounded to 73. Hence, we regard D_I as binomial: $(n, p) = (73, 0.837)$.

Given D_I, the daily total in the initial interval, we let these arrivals be i.i.d. over the initial interval [8:00, 12:20], with a pdf proportional to the continuous two-piece arrival rate function in Figure 8, i.e., with a pdf equal to the arrival rate function divided by its integral over the interval.

In Kim et al. (2015b), binomial-uniform and Gaussian-uniform models were proposed. Our model differs in two respects. First, we treat the outside interval [12:30, 13:00] separately. Second, we treat the initial interval similarly, but our more careful analysis suggests a nonuniform density for the individual arrivals. We propose a scaled version of the continuous piecewise-linear curve on the right in Figure 8, which should better fit the actual arrival rate.

6. Guide to Understanding Appointment-Generated Arrival Processes

While diverse appointment systems should have much in common, there also can be important differences. A useful first step when considering appointment systems and appointment-generated arrival processes is to classify the system. Our analysis of the clinic, summarized in Table 6, helps show how that can be done. There are three main steps.

Table 6. Steps to Classify an Appointment-Generated Arrival Process and the Steps' Application to the Arrivals for the Doctor at the Clinic

Category	Issue	For the doctor at the endocrinology outpatient clinic
General	Time frame for arrivals	One morning shift on a single day
	Time from scheduling to appointment	Mostly 1–4 months
	Time sensitivity (urgency) of appointment	Not known
	Repeat vs. new	78% of visits are repeat
	Scale	Moderately large, average daily total of 66
	Variability of the arrival process	Significant but less than Poisson, dispersion $V/M = 0.3$ for daily totals
Schedule	Variability of the schedule	Significant but less than Poisson, dispersion $V/M = 0.3$ for daily totals
	Master schedule	Identifiable as 22 10-minute intervals with batches of size 3
	Primary deviation from the framework	Extra arrivals scheduled outside the main interval
	High or low demand	High demand
	Extent of overload	Overload producing 10% of daily totals
	Manifestation of overload	Overload occurs outside, usually after, the main interval
	Distribution of the main schedule	The data support i.i.d. batches with mean 2.76 in all time slots
Adherence	No-shows	Relatively few no-shows, or about 8.5%
	Unscheduled arrivals	Relatively few unscheduled arrivals, or about two per day (3%)
	Deviations (lateness or earliness)	Significant deviations of about 60 minutes, but mostly early; about 15% late, with average conditional lateness of about 20 minutes

Step 1. *General Classification.* We first identify the *time frame*, which we take to be a day. However, there are two different perspectives: first, the times when the arrivals occur, and second, the times when the appointments are actually made. We primarily focus on the times when the arrivals occur, aiming to understand variability over the day.

However, as in the clinic studied here, the appointments may have been made over a much longer time frame, weeks or even months before the appointment day, so the delay in getting an appointment may occur over a longer time scale. With such long delays between the date that the appointment is scheduled and the actual appointment date, we have observed that it is important to consider whether arrivals represent, perhaps routinely, repeat visits or new requests. Especially in healthcare, an important question is whether the system can respond well to urgent requests for service. Time sensitivity or urgency was not part of the clinic arrival data analyzed here, but we were able to identify repeat visits, which accounted for 78% of all visits. It is important to recognize that long times between appointments being made and actual clinic visits for those appointments do not necessarily mean that patients with urgent problems are experiencing excessive delays before their needs can be addressed. In general, for healthcare appointment systems, it would be useful to have information on the *delay sensitivity or urgency* of the service to be provided.

We next focus on the *scale*, determined by the typical daily totals. Is the scale large or small? The clinic doctors considered here operate on a large scale, with our specific doctor seeing about 66 patients in each shift (A.M. or P.M.).

Assuming that our goal is to understand the arrival process over a single day and possibly to make improvements in this process, the next question is the *level of variability in the appointment-generated arrivals*. Are the arrivals highly regular or not? The analysis is devoted to the case in which the arrivals exhibit significant variability. An initial rough classification of the variability is the *dispersion* or variance-to-mean ratio V/M of the daily totals.

The remaining classification is aimed at exposing the primary sources of the variability observed in the arrivals. Careful analysis is then devoted to identifying and quantifying the important sources of that variability. Here, it is natural to start with the schedule.

Step 2. *The Schedule.* Given that the actual arrivals are irregular, we ask if the scheduled arrivals are also irregular, exhibiting a significant additional level of variability. For our doctor, we found that the schedule is indeed quite irregular, exhibiting significant variability, and that too can be roughly quantified based on the dispersion of the daily totals. In fact, we concluded that the primary source of variability in the arrivals is the variability in the schedule. This is supported by the fact that both the variance and the dispersion of the scheduled daily totals are greater than for the actual arrivals.

Whether the schedule is regular, we want to identify the master schedule, if possible. In general, a first step in analyzing the schedule is to infer this framework. An orderly framework might be communicated by system managers, but it is important to consider data showing what actually happened. From our examination of the schedule for the 22 A.M. shifts, we were able to identify a stationary framework involving small batches of arrivals at 10-minute intervals during the time interval [8:50, 12:20].

We then ask what the major deviations from this framework are. In the present analysis, we found that the batch sizes in each time slot are variable, but the largest deviation from that framework was caused by extra arrivals scheduled outside the main time interval.

In general, it is important to determine whether the service system is a *high-demand* or *low-demand system*. Is the variability the result of an uncertain ability to fill the master schedule in the presence of low demand or of an uncertain response to pressures to meet high demand? Or do we see a combination of these? We concluded that our doctor in the clinic operates as a high-demand system, with a significant response to high demand.

We then come to the distribution of the scheduled arrivals in the main interval. We concluded in Sections 2.5 and 2.6 that the scheduled arrivals in the 22 daily time slots in the main interval can be regarded as i.i.d. random variables with the distribution in (3), which has mean 2.76. We found relatively low variability in the scheduled arrivals within the main interval.

Step 3. *Adherence to the Schedule.* We next shift attention to the adherence to the schedule. Here, we focus on three ways that the arrivals might not adhere to the schedule: (i) no-shows, (ii) extra unscheduled arrivals, and (iii) deviations in actual arrival times from the scheduled times. Since our clinic data included cancellations, no-shows were easily identifiable as scheduled arrivals that never occurred. Given that all arrivals were included in our clinic data and that our definition of the schedule was based on its value at the end of the previous day, we defined unscheduled arrivals as arrivals that were scheduled and arrived on the current day.

It is well known that no-shows and unscheduled arrivals can be quite frequent in appointment-generated arrival processes. However, in the clinic studied here, there were relatively low percentages of no-shows and unscheduled arrivals. In particular, the average percentage of no-shows for our doctor was about 8.5%. This level was fairly constant over the day but was somewhat higher during the first intervals of the A.M. shift. The average number of unscheduled arrivals in the clinic was only about two per day, which was 3% of the daily total. About half of those occurred outside the main interval, again indicating an effort by the clinic to respond to high demand.

We observed that the actual arrival times deviated significantly from the scheduled arrival times, with an average earliness of 45 minutes and an average lateness of 21 minutes. The deviations were most were due to early arrivals; only about 6% of the arrivals were late by more than 15 minutes. Overall, we conclude that the adherence to the schedule was good relative to that in other appointment systems.

7. Conclusions

The Principal Source of Variability Is the Schedule. In this paper, we have examined an appointment-generated arrival process for one doctor in an endocrinology clinic. As a consequence of the appointment system, the arrival process tends to be much less variable than a Poisson process, but it is also not nearly a regular deterministic arrival process. The dispersion (variance-to-mean ratio) is about 0.3. As others have observed before, some variability is a result of no-shows, extra unscheduled arrivals, and deviations of the actual arrival times from the scheduled appointment times, but Section 3.3 shows that the dominant source of variability in the arrival process is the schedule itself. In particular, surprisingly, the inequality in (16) shows that the dispersion of the daily schedule is actually greater than the dispersion of the daily arrivals itself.

New Stochastic Arrival Process Models. Our data analysis has culminated in both a detailed stochastic model in Sections 2.8 and 3 and a simplified stochastic model in Section 5.3 that can be used to simulate the arrival process of patients to see the doctor in the clinic. The fitting process should be useful for analyzing the other doctors in this clinic as well as for other applications, and simulation experiments can be used to evaluate operational procedures in the clinic.

What Is Generalizable? (i) Variations of the specific arrival process stochastic models developed here may be useful for analyzing other outpatient clinics, but what we think is widely generalizable is *the data-analysis process, rather than the model*. Consistent with earlier work, we advocate carefully examining no-shows, extra unscheduled arrivals, and punctuality. However, before taking those steps, we recommend looking at randomness in the schedule. It may even be important to view the schedule as a stochastic process. We do not have data on the original demand in the current analysis, but we would also advocate collecting information on requests for appointments, including ones that were not scheduled or that were moved to alternate days and times. Additionally, we recommend determining how the schedule relates to the original demand.

(ii) The specific arrival process models may also be useful more widely. Especially promising is the parsimonious model with Gaussian daily totals and, given those daily totals, i.i.d. arrival times within the day with a nonuniform probability density that takes account of the earliness and lateness of the patients. It is reasonable to anticipate that the earliness or lateness will alter the arrival rate during the day, as we have discovered.

(iii) Even more broadly, it is important to recognize that appointment-generated arrival processes are likely to be neither solely deterministic and evenly spaced nor solely Poisson; rather, many systems will have variability in between those two extremes, just as we have seen.

What Is Missing? While we think that our concentrated focus on the clinic arrival process can help in modeling appointment-generated arrival processes, we have not discussed the service-provisioning context beyond our general description of network structure in Section 1.3. In particular, we do not have data on the patient departure times or the level of congestion in the clinic. It was our sense that the clinic was well run, without major operating problems, and not operating in a "heavy-traffic" regime, but we did not have supporting data. To properly understand an arrival process in a queueing application, we think that it is important to understand the full context as well as to look at the arrival process data. We have emphasized a broader context by our focus on the schedule as well as the time of arrival, but other important context might not be included.

What Is the Practical Relevance? In this paper, we have not conducted a complete performance analysis of the endocrinology outpatient clinic, so we have not yet improved the performance of that clinic. However, based on the long history of modeling and analysis of outpatient clinics discussed in Section 1, modeling and analysis can improve system performance. Thus, we did this work with the conviction that improved arrival process models can produce improved performance.

We see two principal ways that the stochastic model of the appointment system can be used to improve the performance of the clinic, and similar stochastic models can also be used to improve performance in other appointment system applications. First, the model provides a basis for analyzing the performance of the clinic with the given arrival process by conducting standard performance (queueing) analyses after incorporating an additional detailed analysis of the patient processing and flow after arrival, which we do not consider here. Second, the model can be used to consider alternative scheduling strategies to achieve various objectives, such as reducing the variability of the schedule and thus reducing the variability in the doctor workloads or ensuring that patients with urgent needs have limited delays in getting an appointment.

Classification of Appointment-Generated Arrival Processes. In addition to gaining a better understanding of the appointment-generated arrival process in the endocrinology clinic, we have learned how to think about appointment-generated arrival processes more generally. A useful first step when considering appointment systems and appointment-generated arrival processes is to classify the system, as we did in Table 6 for our analysis of the clinic. For any new appointment system to be considered, we recommend seeking this information. After evaluating both the schedule and adherence to the schedule by comparing them to what is desired, one could consider ways to improve both the schedule and the adherence.

Acknowledgments

The authors thank Dr. Sang-Man Jin in the Division of Endocrinology and Metabolism at Samsung Medical Center for providing advice, Ponni Vel for her help in the initial stages of this study, and Mohamad Soltani for a helpful comment.

References

Araman VF, Glynn PW (2012) Fractional Brownian motion with $H < 1/2$ as a limit of scheduled traffic. *J. Appl. Probab.* 49(3):710–718.

Bailey NTJ (1952) A study of queues and appointment systems in hospital outpatient departments, with special reference to waiting times. *J. Roy. Statist. Soc.* 14(2):185–199.

Cayirli T, Veral E (2003) Outpatient scheduling in health care: A review of literature. *Production Oper. Management* 12(4): 519–549.

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

Chand S, Moskowitz H, Norris JB, Shade S, Willis DR (2009) Improving patient flow at an outpatient clinic: Study of sources of variability and improvement factors. *Health Care Management Sci.* 12(3):325–340.

Feldman J, Liu N, Topaloglu H, Ziya S (2014) Appointment scheduling under patient preference and no-show behavior. *Oper. Res.* 62(4):794–811.

Fetter RB, Thompson JD (1965) The simulation of hospital systems. *Oper. Res.* 13(5):689–711.

Green LV, Savin S, Murray M (2007) Providing timely access to care: What is the right patient panel size? *Joint Commission J. Quality Patient Safety* 33(4):211–218.

Guo M, Wagner M, West C (2004) Outpatient scheduling—A simulation approach. Ingalls RG, Rossetti MD, Smith JS, Petyers BA, eds. *Proc.* 2004 *Winter Simulation Conf., Washington, DC*, 1981–1987.

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

Hall RW, ed. (2006) *Patient Flow: Reducing Delay in Healthcare* (Springer, New York).

Hall RW, ed. (2012) *Handbook of Healthcare System Scheduling* (Springer, New York).

Harper PR, Gamlin HM (2003) Reduced outpatient waiting times with improved appointment scheduling: A simulation modelling approach. *OR Spectrum* 25(3):207–222.

Honnappa H, Jain R, Ward AR (2015) A queueing model with independent arrivals, and its fluid and diffusion limits. *Queueing System* 80(1–2):71–103.

Hassin R, Mendel S (2008) Scheduling arrivals to queues: A single-server model with no-shows. *Management Sci.* 54(3): 565–572.

Jacobson SH, Hall SN, Swisher JR (2006) Discrete-event simulation of health care systems. Hall RW, ed. *Patient Flow: Reducing Delay in Healthcase Delivery*, Chap. 8 (Springer, New York), 211–252.

Jouini O, Benjaafar S (2012) Queueing systems with appointment-driven arrivals, non-punctual customers, and no-shows. Working paper, École Centrale Paris, Châtenay-Malabry, France.

Jun JB, Jacobson SH, Swisher JR (1999) Application of discrete-event simulation in health care clinics: A survey. *J. Oper. Res. Soc.* 50(2):109–123.

Kaandorp GC, Koole G (2007) Optimal outpatient appointment scheduling. *Health Care Management Sci.* 10(3):217–229.

Kim S-H, Vel P, Whitt W, Cha WC (2015a) Analysis of arrival data from an endocrinology clinic. Columbia University, http://www.columbia.edu/~ww2040/ApptAppendix011715.pdf.

Kim S-H, Vel P, Whitt W, Cha WC (2015b) Poisson and non-Poisson properties in appointment-generated arrival processes: The case of an endocrinology clinic. *Oper. Res. Lett.* 43(3):247–253.

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

Liu N, van de Ven PM, Zhang B (2016) Managing appointment scheduling under patient choices. Working paper, Boston College, Boston, MA.

Liu N, Ziya S, Kulkarni VG (2010) Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manufacturing Service Oper. Management* 12(2):347–364.

Luo J, Kulkarni VG, Ziya S (2012) Appointment scheduling under patient no-shows and service interruptions. *Manufacturing Service Oper. Management* 14(4):670–684.

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

Neal RD, Lawlor DA, Allgar V, Colledge M, Ali S, Hassey A, Portz C, Wilson A (2001) Missed appointments in general practice: Retrospective data analysis from four practices. *British J. General Practice* 51(471):830–832.

Ross SM (1996) *Stochastic Processes*, 2nd ed. (Wiley, New York).

Segal M, Whitt W (1989) A queueing network analyzer for manufacturing. Bonatti M, ed. *Teletraffic Sci. New Cost-Effective Systems, Networks Services Proc.: ITC* 12, *Proc.* 12*th Internat. Teletraffic Congress* (Elsevier, North-Holland), 1146–1152.

Swartzman G (1970) The patient arrival process in hospitals: Statistical analysis. *Health Services Res.* 5(4):320–329.

Swisher JR, Jun JB, Jacobson SH, Balci O (2001) Modeling and analyzing a physician clinic environment using discrete-event (visual) simulation. *Comput. Oper. Res.* 28(2):105–125.

Wang R, Jouini O, Benjaafar S (2014) Service systems with finite and heterogeneous customer arrivals. *Manufacturing Service Oper. Management* 16(3):365–380.

Welch JD, Bailey NTJ (1952) Appointment systems in hospital outpatient departments. *The Lancet* 259(6718):1105–1108.

Whitt W (1982) Approximating a point process by a renewal process, I: Two basic methods. *Oper. Res.* 30(1):125–147.

Whitt W (1983) The queueing network analyzer. *Bell Laboratories Technical J.* 62(9):2779–2815.

Zacharias C, Armony M (2017) Joint panel sizing and appointment scheduling in outpatient care. *Management Sci.* Forthcoming.

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