IIE Transactions on Healthcare Systems Engineering (2013) **3**, 263–279

ISSN: 1948-8300 print / 1948-8319 online
DOI: 10.1080/19488300.2013.858379

Guidelines for scheduling in primary care under different patient types and stochastic nurse and provider service times

HYUN-JUNG OH[1], ANA MURIEL[1], HARI BALASUBRAMANIAN[1,*], KATHERINE ATKINSON[2] and THOMAS PTASZKIEWICZ[2]

[1]*University of Massachusetts, Amherst, 160 Goverenors Drive, Amherst, MA 01003,USA*
E-mail: hbalasubraman@ecs.umass.edu
[2]*Atkinson Family Practice, Amherst, MA, USA*

Received May 2013 and accepted October 2013

Scheduling in primary care is challenging because of the diversity of patient cases (acute versus chronic), mix of appointments (pre-scheduled versus same-day), and uncertain time spent with providers and non-provider staff (nurses/medical assistants). In this paper, we present an empirically driven stochastic integer programming model that schedules and sequences patient appointments during a work day session. The objective is to minimize a weighted measure of provider idle time and patient wait time. Key model features include: an empirically based classification scheme to accommodate different chronic and acute conditions seen in a primary care practice; adequate coordination of patient time with a nurse and a provider; and strategies for introducing slack in the schedule to counter the effects of variability in service time with providers and nurses. In our computational experiments we characterize, for each patient type in our classification, where empty slots should be positioned in the schedule to reduce waiting time. Our results also demonstrate that the optimal start times for a variety of patient-centered heuristic sequences consistently follow a pattern that results in easy to implement guidelines. Moreover, these heuristic sequences and appointment times perform significantly better than the practice's schedule. Finally, we also compare schedules suggested by our two-service-stage model (nurse and provider) with those that only consider the provider stage and find that the performance of the provider-only model is 21% worse than that of the two-service-stage model.

1. Introduction

Primary care providers are typically the first point of contact between patients and health systems. They include family physicians, general internists, and pediatricians. Compared to specialty practices, family-focused primary care involves a higher variety of cases: the same team cares for patients of all ages, from birth to end of life, who suffer from various types of ailments related to both their physical and mental health. There are multiple dimensions to variability in primary care: nature of patient complaint (acute versus chronic); mix of appointments (pre-scheduled versus same-day); and time spent with providers and non-provider staffs (nurses/medical assistants). This variability in turn influences patient wait time and the utilization of providers. Scheduling in the face of such variability is a significant challenge for practices.

In this paper, we present an empirically driven stochastic integer programming model that schedules and sequences patient appointments during a work day session. The objective of this model is to minimize a weighted measure of provider idle time and patient wait time. Key features of the model include: patient classification to accommodate different chronic and acute conditions seen in the practice; adequate coordination of patient time with a nurse and a provider; and strategies for introducing slack in the schedule to counter the effects of variability in service time with providers and nurses. While outpatient scheduling is a well-studied topic (see Cayirli and Veral, 2003 and Gupta and Denton, 2008, for a review), the current article brings together the disparate elements mentioned above—which have typically been studied only in isolation—into a single, tractable optimization framework. For example, many stochastic optimization approaches schedule only the provider (Robinson and Chen, 2003, and Denton and Gupta, 2003), but they do not coordinate patient service time with the nurse or consider the diversity of patient conditions. We use the model to create broader guidelines that can help practices carry out more effective scheduling while staying sensitive to current protocols and operational constraints. We also compare the proposed schedules to actual schedules used in practice.

*Corresponding author

Table 1. Appointment mix

	Number	Percentage
Total number of scheduled patients*	420	
Total number of observed patients	364	
No-shows	13	3
Cancellations and reschedules	19	5
Prescheduled appointments	317	75
Same-day appointments	103	25
30-min. appointments	121	33
15-min. appointments	243	67

*Total number of scheduled patients includes patients who were scheduled during the course of the study, including no-shows, cancellations and reschedules, and those who only received nurse care.

Our model is motivated and calibrated using empirical data collected at a three-provider family medicine practice in Massachusetts. The data was collected using a time-motion study conducted on nine work days in summer and fall of 2011. All told, 420 patients were scheduled for appointments during the course of the research study. We observed 364 patients from beginning to end. The total numbers of patients who were scheduled and patients who were actually observed are different because some patients saw only a nurse to receive a flu shot or simple treatment. A descriptive summary of the data is provided in Table 1. The practice schedules patients in 15-min. increments and reserves either a 15- or 30-min. appointment slot for each patient depending on their predicted complexity. Same-day appointments are allocated a 15-min. slot, and occasionally double-booked. We analyze the data collected in more detail in Section 3.

While the results of our time study and the models we tested may seem restricted to the practice we worked with, they are in fact fairly general. This is because the majority of the primary care practices in the United States are small and follow similar scheduling processes. In fact, 32% of the practices in the U.S. are solo practices, and 78% of the practices consist of five providers or less (Bodenheimer and Pham, 2010). Moreover, the types of patient conditions seen at this practice—chronic conditions such as diabetes, depression, fatigue, routine physicals for adults and children; and acute conditions such as sore throat and migraine—are representative of the patients seen in all primary care practices. We provide greater detail on the appointment durations for these conditions in Section 3.

The rest of the article is structured as follows. In Section 2, we review the literature most relevant to the problem. In Section 3, we analyze the empirical data and motivate the scheduling and sequencing model discussed in Section 4. In Section 5, we use this mathematical model to address five focused questions relevant for the practice. We summarize our conclusions and implications for practice in Section 6.

2. Literature review

Research on outpatient appointment scheduling is well established and growing. A comprehensive review of the topic is provided in Cayirli and Veral (2003) and Gupta and Denton (2008). Cayirli and Veral (2003) classify analysis methodologies into queuing theory, mathematical programming methods, and simulation studies. Among these methodologies, we use mathematical programming methods, driven by empirical data, since they have benefits distinct from other methodologies: unlike queuing theory, no particular assumptions are necessary such as the distributions of inter-arrival times, distributions of service times, and queue capacity; and unlike simulation studies, we can find the exact optimal solution rather than using only heuristics (Berg 2012). To maintain consistency with the literature, we use the broader term outpatient practice instead of primary care practice in this section. In addition, it is worth to clarify definitions among a scheduling rule, a sequencing rule and an appointment rule; the scheduling rule is composed of the sequencing rule that determines the sequence in which patients will be seen and the appointment rule which assigns specific appointment times to these patients.

Our goal is to provide easy-to-implement scheduling guidelines for primary care practices using a stochastic integer programming approach. We therefore review literature relevant to two issues: scheduling guidelines irrespective of the methodology, and mathematical programming approaches to appointment scheduling. We consider any application setting in outpatient healthcare delivery, including surgery.

The most well-known outpatient appointment rule is the *Bailey-Welch* rule (Bailey, 1952; Welch, 1964) which assigns two appointments for the very first slot and one appointment in the rest of the slots. This rule was shown using queuing models and simulation studies with mean service times. Ho and Lau (1992) using simulation also prove that the Bailey-Welch rule is robust. Soriano (1966) compares one appointment per slot using mean service times to the *Two-at-a-time* rule (two double-booked appointments, followed by an empty slot) using queuing theory. He finds that Two-at-a-time is successfully applied to an outpatient department in significantly reducing wait time.

Kaandorp and Koole (2007) use a heuristic local search algorithm to optimize wait time, idle time, and overtime with homogeneous patients with equal slot lengths. They consider three parameters: probability of no-shows, average service time, and total number of patients. They conclude that dome-shaped inter-appointment times are robust; dome-shaped indicates that inter-appointment intervals first increase and then decrease. The optimal appointment rule is very similar to the Bailey-Welch rule with particular parameter values (weights). Using a simulation-based optimization, Klassen and Yoogalingam (2009) find

times, *plateau-dome* pattern (slot lengths in the dome part are equal), is robust in considering various environment factors, such as number of appointment slots, probability of no-shows, and session lengths.

The literature cited above assumes fixed inter-appointment times. Chew (2011) relaxes this assumption and focuses on determining inter-appointment times given a known number of slots from historical data using a simulation-based heuristic algorithm to minimize expected wait time, idle time, and overtime. He finds that as the unit cost for wait time is higher, the inter-appointment times are increased; as the unit cost for idle time is higher, the inter-appointment times are decreased; and if the unit cost for overtime is increased, the last slot is long enough to prevent overtime.

Hassin and Mendel (2008) study the two types of appointment systems, non-fixed inter-appointment times and fixed inter-appointment times, by using queuing systems with a single server considering different show rates for each patient. They find that with no-show rates, their optimal schedule with non-fixed inter-appointment times seems dome-shaped since the appointment interval increases for the first few appointments, then stays almost the same, and then decreases for the last few appointments. With fixed inter-appointments, the slot length decreases as no-show rates increases.

The papers discussed above assume patients to be homogenous. However, the outpatient practice generally consists of various patient types, each of whose service times involves significantly high variability. Klassen and Rohleder (1996) evaluate different scheduling rules with different types of patient and equal slot lengths by conducting simulation. They conclude the best sequencing rule is to allocate all low variance patients at the beginning of the session and high variance patients toward the end to strike a balance between wait time and idle time. Although this sequencing rule is practical, it is often difficult to have knowledge of variance of each patient type. Based on our empirical study, we find that patients differ in their *mean* durations, but we are unable to classify patients by variance since all patient types vary significantly in their appointment durations (see Section 3). Hence, mean service durations could be more tractable to use in patient classification. Cayirli *et al.* (2006) employed mean service times to classify two different patient types (new and return patients, which correspond to long and short mean service times, respectively). They use discrete event simulation to evaluate various types of scheduling rules using empirical data with the goal of reducing wait time and idle/overtime. It is interesting to note that although service time variability is statistically different, it has less significant impact on the performance in comparison to the clinic size, no-shows, walk-ins, and patient punctuality. Among appointment rules, the Bailey-Welch rule is close enough to the efficient frontiers and can be applied to all sequencing rules they tested. Cayirli *et al.* (2008) extends the study of Cayirli *et al.* (2006) by comparing schedules with equal inter-appointment times with schedules that set two different inter-appointment times equal to the mean of new and return patient service times, respectively. They show that when the cost of provider idle time is high relative to that of patient wait, a schedule using an SPT (shortest processing time) sequence and following the Bailey-Welch rule along with the two different inter-appointment lengths performs very well.

Most papers have considered only a single step, the provider step, in the patient flow process. However, Gul *et al.* (2011) do consider three independent steps, intake, procedure, and recovery steps, in outpatient procedure centers with the goal of minimizing the expected patient wait time and overtime. They first use discrete event simulation; then they develop a genetic algorithm (GA) (Holland, 1975) to analyze simple sequencing heuristics. Among heuristics, SPT performs the best. In addition, they use their GA to see the impact of rescheduling procedures within a given time-horizon of n-days. They conclude that the rescheduling procedures significantly help reduce wait and overtime since a procedure can be assigned to a lower utilization day.

We next turn to papers that use stochastic linear programming. Outpatient surgical scheduling is relevant to our work since procedure durations—like service times in primary care—are highly variable. The most relevant papers are Robinson and Chen (2003), Denton and Gupta (2003), Denton *et al*. (2007), Mancilla and Storer (2012), Berg (2012) and Saremi *et al*. (2013).

Robinson and Chen (2003) formulate a stochastic linear program with empirically determined distributions of surgery service times in order to determine inter-appointment times given the known patient sequences. The objective is to minimize the expected weighted sum of patient wait time and provider idle time. They solve it by using Monte Carlo integration (see Hammersley and Handscomb, 1964; Halton, 1970; and Fishman, 1996). They propose a scheduling rule using two different inter-appointment durations, one is applied to the first appointment and the other is assigned to the remaining appointments, and show that it is close to the optimum. Denton and Gupta (2003) also optimize inter-appointment times by formulating a two-stage stochastic linear program, considering different coefficients of wait, idle and overtime. They exploit the L-shaped algorithm (Van Slyke and Wets, 1969) with sequential bounding. They show that inter-appointment times display a dome shape when the ratio of idle to wait cost is high, while look more uniform when the cost ratio is low.

Since the surgeries were all of the same type, the above papers focus only on optimal appointment times. Denton *et al*. (2007) and Mancilla and Storer (2012) consider appointment times as well as the sequencing decisions for different surgery types. Denton *et al*. (2007) optimize the sequences and appointment times of surgeries in operating rooms using a two-stage stochastic programming model. The surgery duration and the schedule are derived from

historical data. They find that it is hard to review all possible combination of sequences ($n!$) in their stochastic programming formulation. Thus, they compare actual schedules used in the practice with three different heuristics. Their results confirm that low-variance surgeries sequenced earlier in the schedule provides robust performance. In addition, Mancilla and Storer (2012) expand the work of Denton *et al*. (2007). They develop new algorithms using Bender's decomposition to determine the optimal appointment times in settings with fixed slot lengths. In Denton *et al*. (2007), appointment time decisions are not restricted by fixed slot lengths. Mancilla and Storer (2012) compare the cases with equal vs. unequal costs for the different surgeries. In the case of equal costs, the sequencing rule by Denton *et al*. (2007), the assignment of shorter variance cases first, performs quite well. In the case of unequal costs, however, the algorithms based on Bender's decomposition outperform the shorter variance first assignment.

Some papers not only use a mathematical model to find the optimal scheduling rules but also implement them in simulation studies to measure performance. Berg (2012) determine optimal scheduling rules and booking number of procedures using a two-stage stochastic mixed integer program with a single server and five different types of procedures in outpatient centers. They consider no-show rates; an attendance binary random variable is defined by the no-show probability. They employ two decomposition methods based on the classic L-shaped method and a progressive hedging heuristic (Rockafellar and Wets, 1991). Each method improves solution times and optimality gaps. Their findings are the following: patients who have high variance procedure durations or high no-show probability need to be scheduled towards the end of the session; a double booking occurs as no-show probability increases; the Bailey-Welch rule is followed in the optimal schedule; and the optimal number of patients to schedule is quite robust with regard to estimates of the fixed cost of running the suite. In addition, they use discrete event simulation to compare the actual sequences and schedules from the practice with solutions derived by their single server stochastic model. The patient flow structure in the simulation is similar to Gul *et al.* (2011) and models the registration step and three types of procedure rooms. The stochastic program solutions yield up to 63% higher expected profits than the actual one followed by the practice.

Saremi *et al*. (2013) propose a simulation-based tabu search and two possible enhancements using integer and binary programming methods. The models consider different types of patients with stochastic durations in a three-step patient flow process: pre-operation, procedure, and recovery. They find that the enhanced searching methods perform significantly better in terms of wait time and computation time and highlight the superior performance of four scheduling rules: (i) dome-shaped according to the mean service time (μ), (ii) dome-shaped according to adjusted service times ($\mu+\alpha\sigma$), (iii) increasing variance, and (iv) increasing coefficient of variability of service time.

Muthuraman and Lawley (2008), Chakraborty *et al*. (2010), Lin *et al.* (2011), Turkcan *et al*. (2011), and Chakraborty *et al*. (2012) focus on scheduling decisions as patient call-ins arrive sequentially in an outpatient practice. Patients are identical as far as service times are concerned, but differ based on their probability of no-show. These papers establish the importance of considering heterogeneous no-show probabilities of patients in appointment scheduling; they also consider the interaction between heterogeneous no-shows and aspects such as impact of pre-defined slot structures and fairness in performance across patients.

In summary, outpatient appointment scheduling is a well studied area. We contribute to the literature in the following ways. Our empirical study provides estimates on service time durations for common patient conditions typically seen in primary care. We then use this data to propose a practical, new patient classification scheme, and use the classification to develop scheduling guidelines. Previous mathematical programming approaches mostly consider a single stochastic service step; if multiple steps are modeled, service times in each step are all assumed to be deterministic. In our stochastic integer program, we explicitly consider both nurse and provider steps in the patient flow process, with stochastic service times in both steps that depend on patient type. Furthermore, in our computational results, we consider a variety of heuristic schedules that accommodate patient time-of-day preferences. We demonstrate that these schedules have a specific structure that makes them easy to implement in practice, while providing a good balance of patient wait and provider idle times.

3. Time study: Data analysis

In this section, we present the empirical study that motivated this article. We first describe the practice under consideration and how the data was collected. The remainder of the section summarizes the data and the insights obtained regarding patient flow and the variability in service time with nurse and provider for different patient types and ailments.

3.1. *Data collection methodology*

We collected data at a three-provider family medicine practice in Massachusetts. Figure 1 illustrates the layout of the practice. The black rectangle indicates the location of the observer who conducted the time-study. We gathered data on nine work days: July 7, 18, 22, August 3, 8, and October 5, 7, 8, 9 in 2011. We observed all patients seen by the providers on these days. At the beginning of the day, we examined the list of prescheduled appointments; at the end of the day, we reviewed the list of all appointments including same-day appointments, no-shows, cancellations, and reschedules. We were thus able to collect the data of all patients during a work day. In other words, our data is not merely a sample; it can construct *a complete chronology of patient flow* on the nine work days.

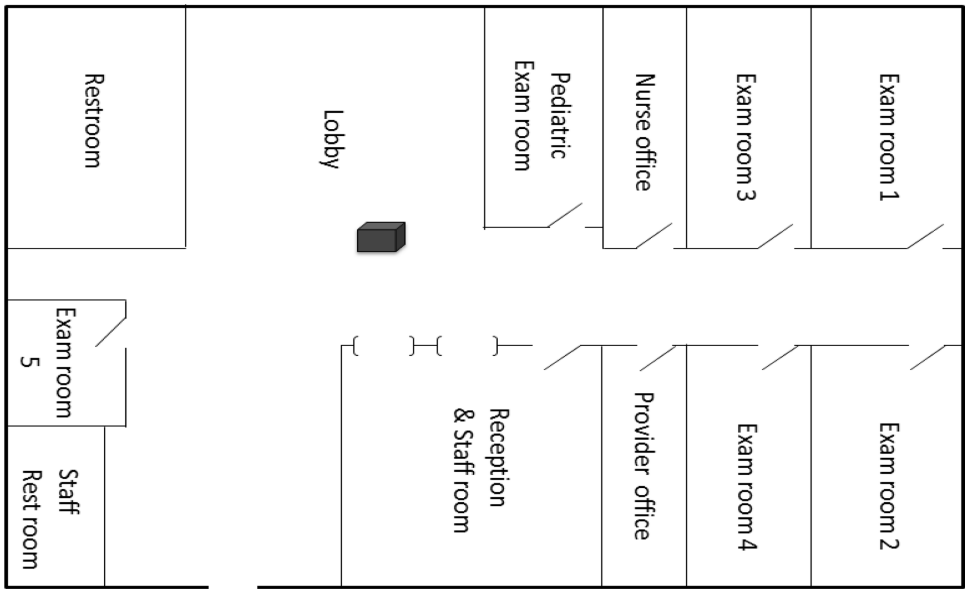

Fig. 1. Layout and observer location at the studied family medicine practice.

Once patients enter the practice, they proceed to the reception desk to notify their arrivals to the receptionist. They wait in the lobby until a nurse calls to examine the patient. After the exam, the nurse flips a flag indicating that the patient can now be seen by the provider. The patient waits in the exam room until a provider is available. Before seeing the patient, the provider flips another flag; and once the appointment has concluded, she flips down all flags. These flags are visible from the lobby, where the observer is present, and allow for the unobtrusive collection of the following time stamps: wait time in the lobby; time with a nurse; wait time in the exam room; time with a provider; and total time of patient visits. In our face and wait time observations, we accounted for the fact that a nurse and/or a provider sometimes returned to visit the patient in the exam room even after the conclusion of the initial service time.

3.2. *Summary of patient flow measures*

Figure 2 presents a box plot, the average and the standard deviation of each indicator of patient flow. On average, patients wait 4 min. in the lobby, spend 12 min. with a nurse,

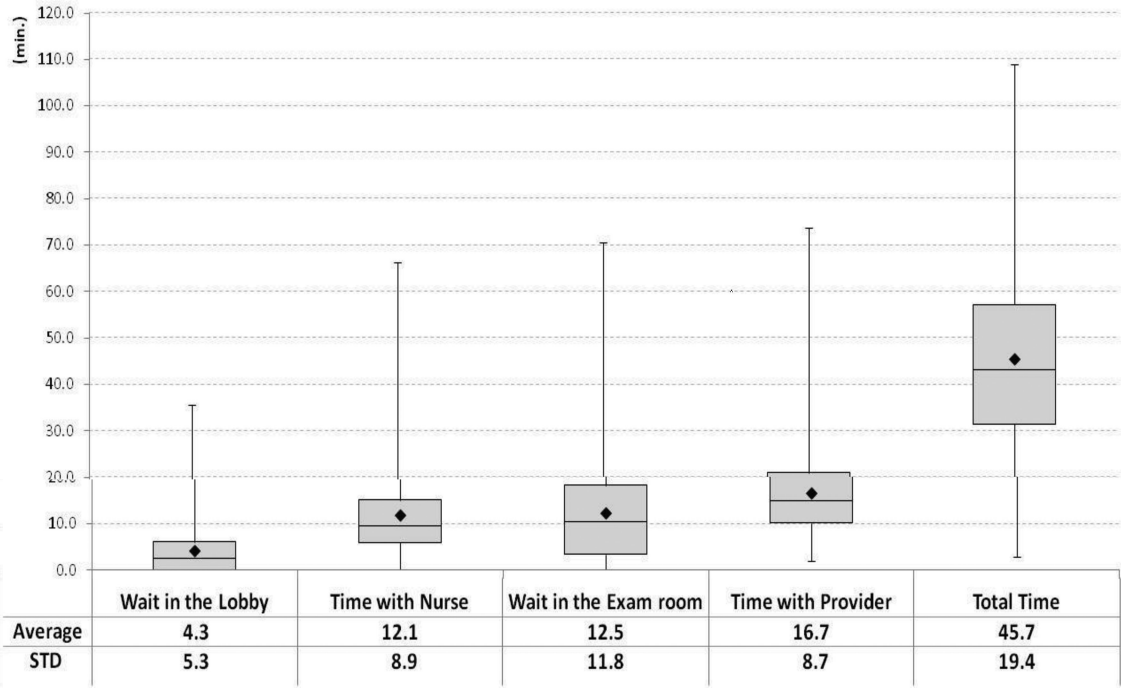

Fig. 2. Box plot of practice performance (min.).

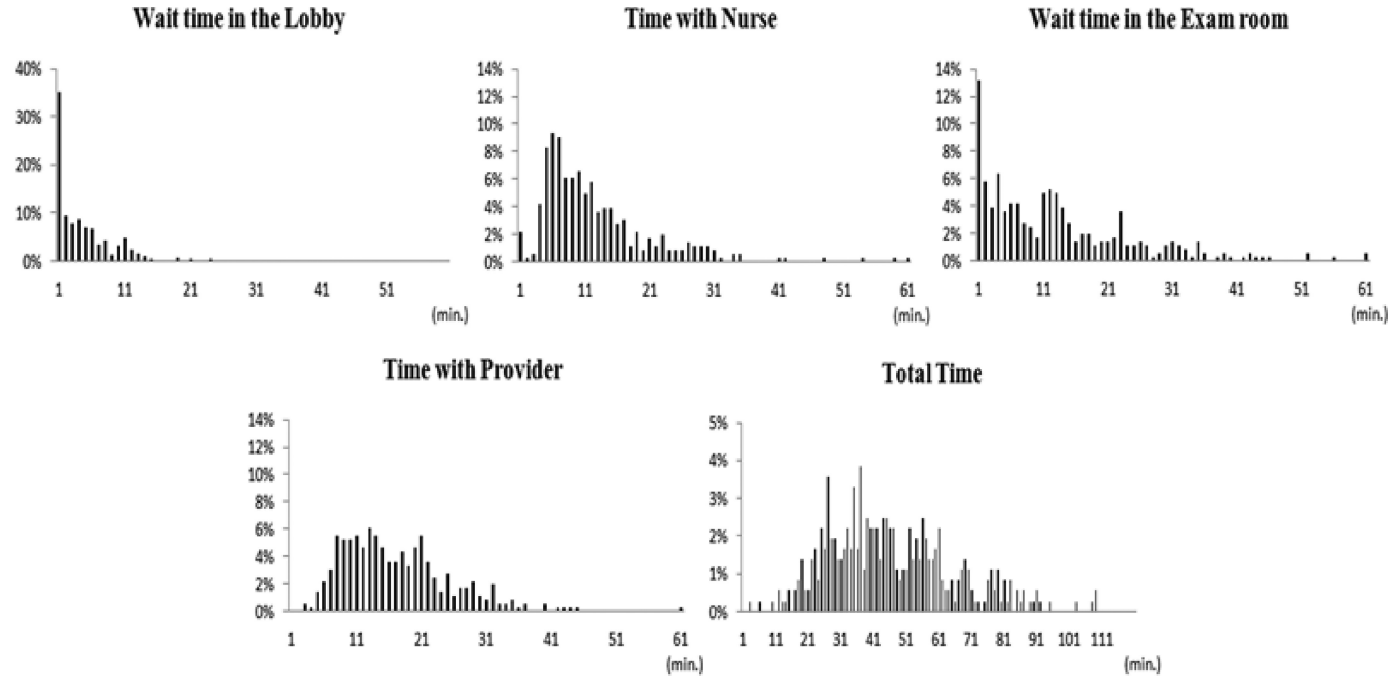

Fig. 3. Distribution of patient flow.

wait 13 min. in the exam room, and finally spend 17 min. with a provider. In total, patients spend 46 min. at the practice. Although at first glance each of the performance indicators appears satisfactory, there is, in fact, significant variability among the time indicators. In particular, wait time in the exam room has a high standard deviation. Distributions of each recorded measure are shown in Figure 3.

Patients must have the necessary amount of service time with nurses and providers. As shown in Figure 3, however, service time with a medical team (nurses and providers) is highly variable; furthermore, the service time distributions for both nurses and providers are skewed to the right. The variability is understandable as these distributions aggregate both 15-min. and 30-min. appointments and a great variety of patient needs. The data shows that 15-min. appointments often exceeded their anticipated durations; in fact, 42% of 15-min. appointments (whether prescheduled or same-day) took longer than 15 min. with providers, and 24% of them exceeded 20 min.

The histograms of both lobby and exam wait times resemble the Exponential distribution. In the lobby, 29% of patients had 0 wait time, and 68% of patients waited fewer than 5 min. In the exam room, we observed that 52% of patients waited more than 10 min. and 10% waited 30 min. or more. These relatively long wait times are of particular concern to the practice, as they erode patient satisfaction. Certainly, waiting in the exam room increases patient discomfort and anxiety, and is not convenient. The distribution of total time that patients spend at the practice, which aggregates all the time measures, understandably looks less skewed.

3.3. *Provider and nurse service time by patient condition*

We found that service times vary significantly depending on the nature of the patient's ailment. Further, as shown in Figure 4, different patient conditions require different amounts of service time with nurses and providers. Notice that while service time with providers is typically higher, nurse times are non-trivial and in some cases higher. For instance, patients scheduled for well-child check-ups or sore throat visits require longer time with nurses because specific medical tests need to be performed. Therefore, coordinating nurse and provider times for the various patient types in the schedule is essential if exam room waiting is to be reduced.

3.4. *Improved appointment classification*

The practice currently schedules two types of appointments, 15-min. and 30-min., in 15-min. slots. In the provider's schedule, a 30-min. appointment takes up two consecutive 15-min. slots. The 30-min. appointments consist of routine physical exams; well-child check-ups; diabetes and chronic condition management; new patient visits; procedures; and migraines and headaches. All other appointments—including same-day requests—are scheduled as 15-min. appointments.

As Table 2 shows, the mean and standard deviation of service times of the patients we observed in our time-study

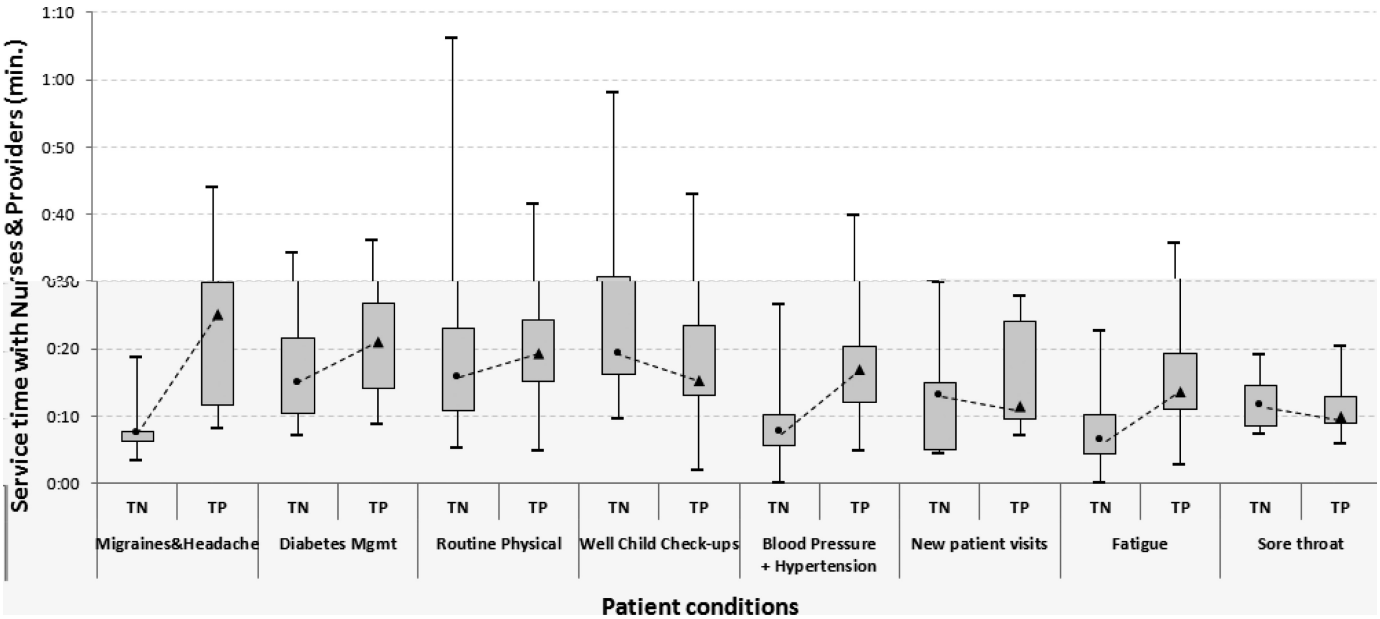

Fig. 4. Box plots of service time with nurses and providers by patient conditions. *TN: time with nurses; TP: time with providers.

are indeed statistically different for these two types of appointments considered by the practice.

Our empirical study suggests that we can further refine the classification of appointments. Based on time requirements, we propose classifying patients into three easy-to-identify groups: prescheduled 30-min. appointments of high complexity (*HC*), which consist of the six conditions mentioned above; prescheduled 15-min. appointments, which include conditions of relatively low complexity (*LC*); and appointments scheduled on short notice, which consist of urgent, same-day appointments (*SD*).

Table 3 shows that the differences among the three groups we propose are indeed statistically significant. The practice currently lumps all LC appointments and SD appointments into the same 15-min. appointment category. On average, however, the LC patient spends three additional minutes compared to the SD patient while still remaining clearly distinct from the HC patient. Note that LC and SD patients will still be scheduled in 15-min. slots. However, if a number of LC appointments are scheduled in succession without slack, wait times are more likely to accumulate than when the same number of SD appointments is scheduled in succession. This subtle point has implications for the scheduling questions we study in the next sections.

The new classification makes also intuitive sense from the point of view of the practice since the SD appointments become known only as the work day progresses, whereas all prescheduled patients, whether in the HC or the LC categories, are known at the beginning of the work day. In addition, SD patients' calls have to be fulfilled at a short notice. The short notice here refers to a few hours or half a day (patients who need immediate care don't fall in this category and are typically directed to an emergency room). Thus, it is important to have slots available towards the end of a session. This also helps reduce the risk of unfilled slots (or double-booking) by allowing the practice to provide a patient who calls in at, say, 8 am with a late morning or early afternoon slot. Indeed, the practice we

Table 2. Service time with nurse and provider by patient type under current patient classification (min.)

	Mean	Standard deviation	*T*-test p-value
30-min.			
Nurse	18.5	10.7	0.000
Provider	19.1	7.9	0.000
15-min.			
Nurse	9.0	5.7	Ref.
Provider	15.6	8.8	Ref.

Table 3. Service time with nurse and provider by patient type under new patient classification (min.)

	Mean	Standard deviations	*T*-test p-value
HC (High Complexity)			
Nurse	17.8	10.7	0.000
Provider	19.5	8.2	0.005
LC (Low Complexity)			
Nurse	8.5	5.1	Ref.
Provider	16.6	9.0	Ref.
SD (Same-day)			
Nurse	9.5	6.1	0.239
Provider	12.7	7.0	0.000

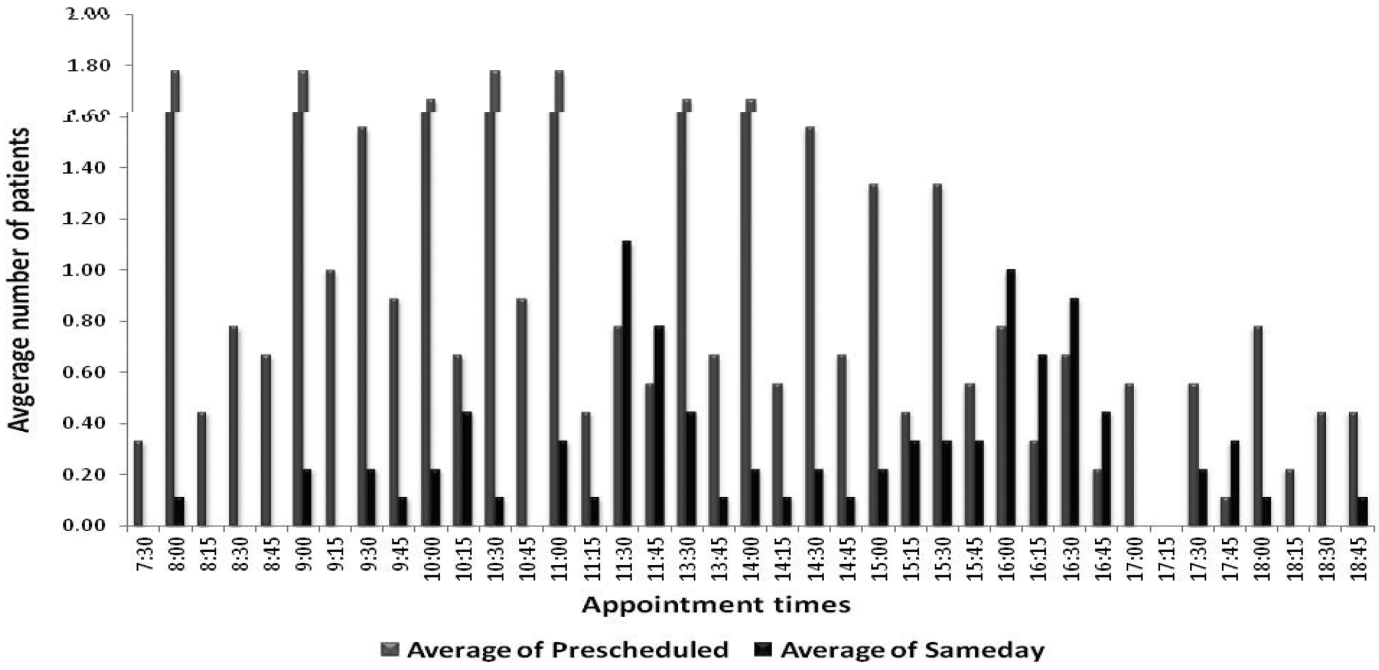

Fig. 5. Pre-scheduled vs. same-day by time of day.

worked with has followed this policy based on our recommendation. See Balasubramanian *et al.* (2013) for a detailed analysis. Figure 5 shows the average number of prescheduled and same-day appointments by time of day for nine work days with two providers working in parallel. SD appointments are mostly scheduled late in the morning. In the afternoon session, however, SD appointments can be more evenly distributed; yet, for the same reasons discussed above, some SD appointments are made available later in the afternoon.

4. Integer programming formulation

4.1. *Model description*

We now present a two-stage stochastic integer program (SIP) for assigning multiple patient types to appointment slots in a *session*. A session refers to a block of time (typically a few hours) either in the morning or afternoon. The morning and afternoon sessions can be decoupled, and their schedules studied independently, since there typically is a break for lunch.

The objective of the SIP is to minimize a weighted measure of provider idle time and patient wait time in the session. Wait time in our model has two components: wait time in the lobby (until the nurse calls), and wait time in the exam room after the nurse exam (until the provider is ready). However, we simply consider total patient wait times as the measure of performance in the computational results. We assume that a provider's calendar for a morning or afternoon session consists of contiguous appointment slots, each having a fixed, predetermined length (15 minutes in our case study). Each patient spends an uncertain amount of time with first the nurse and then the provider. The distribution of these service times depends on the type of patient being scheduled. The number of patients of each type to be scheduled is known beforehand. While this is not the case in reality, we demonstrate in our computational results that the guidelines we develop using our model are robust to changes in the mix of patients scheduled. We also assume that the patients arrive punctually and are not called by the nurse before their scheduled appointment times.

The first-stage decisions of the SIP involve both the *sequence* in which the patient types are scheduled, and the *appointment times* of each patient. Because the slots in our case-study are 15 minutes long, the appointment times are always in 15-min. increments. For any feasible first-stage decisions (which determine the schedule for the session), nurse and provider service times are realized in the second stage, resulting in idle time for the provider and wait time for the patients.

We create 1,000 scenarios or realizations by sampling randomly from the empirical face-time distributions obtained from the field study. We use the sample average approximation method (see Kleywegt *et al.*, 2002).

The two-stage stochastic integer program is described formally below.

Notation:

I	Number of patients to be scheduled in the session, indexed by i, $i = 1,\ldots, I$
S	Number of scenarios, indexed by $s = 1,\ldots, S$

Parameters

α	Weight for idle time
β	Weight for wait time
N_{HC}	Number of patients of type HC to be scheduled
N_{LC}	Number of patients of type LC to be scheduled
N_{SD}	Number of patients of type SD to be scheduled
$\tau_{i,s}^{n,HC}$	Service time with a nurse for patient *i*, if of type *HC,* under scenario *s*
$\tau_{i,s}^{n,LC}$	Service time with a nurse for patient *i* if of type *LC,* under scenario *s*
$\tau_{i,s}^{n,SD}$	Service time with a nurse for patient *i* if of type *SD,* under scenario *s*
$\tau_{i,s}^{p,HC}$	Service time with a provider for patient *i* if of type *HC,* under scenario *s*
$\tau_{i,s}^{p,LC}$	Service time with a provider for patient *i* if of type *LC,* under scenario *s*
$\tau_{i,s}^{p,SD}$	Service time with a provider for patient *i* if of type *SD,* under scenario *s*

Variables

$\tau_{i,s}^{n}$	Service time of patient *i* with a nurse under scenario *s*
$\tau_{i,s}^{p}$	Service time of patient *i* with a provider under scenario *s*
$y_{i,s}^{start}$	Start time of patient *i* with a nurse under scenario *s*
$y_{i,s}^{finish}$	Finish time of patient *i* with a nurse under scenario *s*
$z_{i,s}^{start}$	Start time of patient *i* with a provider under scenario *s*
$z_{i,s}^{finish}$	Finish time of patient *i* with a provider under scenario *s*
$A_i \in \{0, 1\}$	1 if patient *i* is HC, 0 otherwise
$B_i \in \{0, 1\}$	1 if patient *i* is LC, 0 otherwise
$C_i \in \{0, 1\}$	1 if patient *i* is SD, 0 otherwise
X_i	Appointment slot for patient *i, X_i in 0,1,2,. . .,15 for a 4-hour session*

The problem is modeled as the following integer program.

$$Min. \quad \frac{1}{S}\left(\alpha\left[\sum_s\sum_{i=1}^{n}\left(z_{i,s}^{start}-z_{i-1,s}^{finish}\right)\right]+\beta\left[\sum_s\sum_{i=1}^{n}\left(y_{i,s}^{start}-15X_i\right)+\left(z_{i,s}^{start}-y_{i,s}^{finish}\right)\right]\right) \quad (1)$$

Subject to.

$$z_{0,s}^{finish}=0, \quad \forall s\in S \quad (2)$$

$$X_{1,s}=0, \quad \forall s\in S \quad (3)$$

$$\tau_{i,s}^{n}=\tau_{i,s}^{n,HC}\times A_i+\tau_{i,s}^{n,LC}\times B_i+\tau_{i,s}^{n,SD}\times C_i, \quad \forall i\in I, s\in S \quad (4)$$

$$\tau_{i,s}^{p}=\tau_{i,s}^{p,HC}\times A_i+\tau_{i,s}^{p,LC}\times B_i+\tau_{i,s}^{p,SD}\times C_i, \quad \forall i\in I, s\in S \quad (5)$$

$$y_{i,s}^{start}\geq 15X_i, \quad \forall i\in I, s\in S \quad (6)$$

$$y_{i,s}^{start}\geq y_{i-1,s}^{finish}, \quad \forall i\in I, s\in S \quad (7)$$

$$y_{i,s}^{finish}=y_{i,s}^{start}+\tau_{i,s}^{nurse}, \quad \forall i\in I, s\in S \quad (8)$$

$$z_{i,s}^{start}\geq y_{i,s}^{finish}, \quad \forall i\in I, s\in S \quad (9)$$

$$z_{i,s}^{start}\geq z_{i-1,s}^{finish}, \quad \forall i\in I, s\in S \quad (10)$$

$$z_{i,s}^{finish}=z_{i,s}^{start}+\tau_{i,s}^{PCP} \quad \forall i\in I, s\in S \quad (11)$$

$$\sum_{i=1}^{n}A_i=N_{HC} \quad (12)$$

$$\sum_{i=1}^{n}B_i=N_{LC} \quad (13)$$

$$\sum_{i=1}^{n}C_i=N_{SD} \quad (14)$$

$$A_i+B_i+C_i=1 \quad (15)$$

$$A, B, C\in\{0,1\}; \quad y^{start}, y^{finish}, z^{start}, z^{start}\geq 0; \quad X\ int.$$

The objective function (1) minimizes the weighted sum of idle time and wait time over all scenarios. Note that computation of the provider's idle time is based on the difference between the start time of patient *i* and finish time of patient *i*-1. For the patients' wait time in the lobby, we look at the difference between the start time of patient *i* with a nurse and the appointment time. For the wait time in the exam room, we take the difference from the start time of patient *i* with a provider minus the finish time of patient *i* with a nurse. Constraints (2–3) initialize the finish time of the 0th patient with a provider to 0, and the first patient start time with nurse to the beginning of the session, in every scenario. Constraints (4–5) ensure that proper service times are used given the patient type. Constraints (6–8) keep track of start time and finish time of patient *i* with a nurse, as well as set the appointment time given to patient *i*. Constraints (9–11) track start time and finish time of patient *i* with a provider. Constraints (12–14) ensure that the desired number of patients of each type is scheduled in the session. Constraint (15) enforces that only one patient type can be scheduled on the particular slot.

As a benchmark, we also consider a *deterministic* integer program (DIP) by assuming that nurse and provider service times take on their respective average values and have no variability. We use the CPLEX Solver Version 12.4 to solve the SIP and the DIP.

Notice that, for a predetermined patient sequence, the SIP can also be used to optimally determine the appointment times of each patient. The spacing between the scheduled arrivals of two patients determines *slack* in the schedule. Slack prevents the accumulation of patient waiting. Given that sequences can vary from day to day based on patient requests and time-of-day preferences, it is important to derive robust guidelines on where slack should be strategically positioned in the schedule.

4.2. *Calibrating weights in the objective function*

How much should a unit of provider idle time be valued against a unit of patient waiting? This is a recurring issue in all appointment scheduling research (see Robinson and Chen, 2011, for a detailed discussion). We looked at five afternoon schedules from our time-study. The mix of patients varied from one afternoon to another. We compared the schedule used in practice with the schedules generated by the SIP and the DIP. While we tested a wide range of weights, we narrowed down our search to cases where a provider's idle time is equal to or higher than that of patient waiting. This makes intuitive sense since idle time is experienced by a single person while wait time accumulates across all scheduled patients. In addition, high idle time is unacceptable in the primary care practice as it would make it financially unviable.

The results are shown in Figure 6. As an example, DIP 0.8:0.2 implies that the weight on provider idle time is 0.8 and on patient waiting is 0.2 in the DIP.

We observe that the DIP is mostly insensitive to the weights. This is understandable since the DIP does not capture variability and therefore grossly underestimates how wait times accumulate as the day progresses. The SIP, which considers variability, exhibits greater sensitivity toward changes in weights. We notice that the SIP 0.5:0.5 results in much higher idle time than other weight combinations; on average, idle time of the SIP 0.5:0.5 is more than 50 min. in a session with only 10 patients, which would be unacceptable in a primary practice. We also find that the SIP 0.7:0.3 schedules provide low wait times but more idle time than acceptable for the practice, while the SIP 0.9:0.1 schedules provide very little slack and thereby increase patient waiting beyond the desired levels. The practice needs to strike a careful balance between inducing high levels of provider idle time by adding too much slack in the schedule, and observing lengthy patient waits when not adding enough. Fortunately, the SIP 0.8:0.2 schedules tested provide the right balance between these two cases.

These observations are further illustrated in the schedules generated by the SIP when all patients are of the same type (the homogeneous patient case). Notice in Figure 7 (a) and (b) that in the 0.8:0.2 schedule, the number of empty slots (slack) is exactly one fewer and one more than the 0.7:0.3 and 0.9:0.1 schedules, thus striking a balance. Consequently, we will use the 0.8:0.2 weight combination in the remainder of our computational experiments.

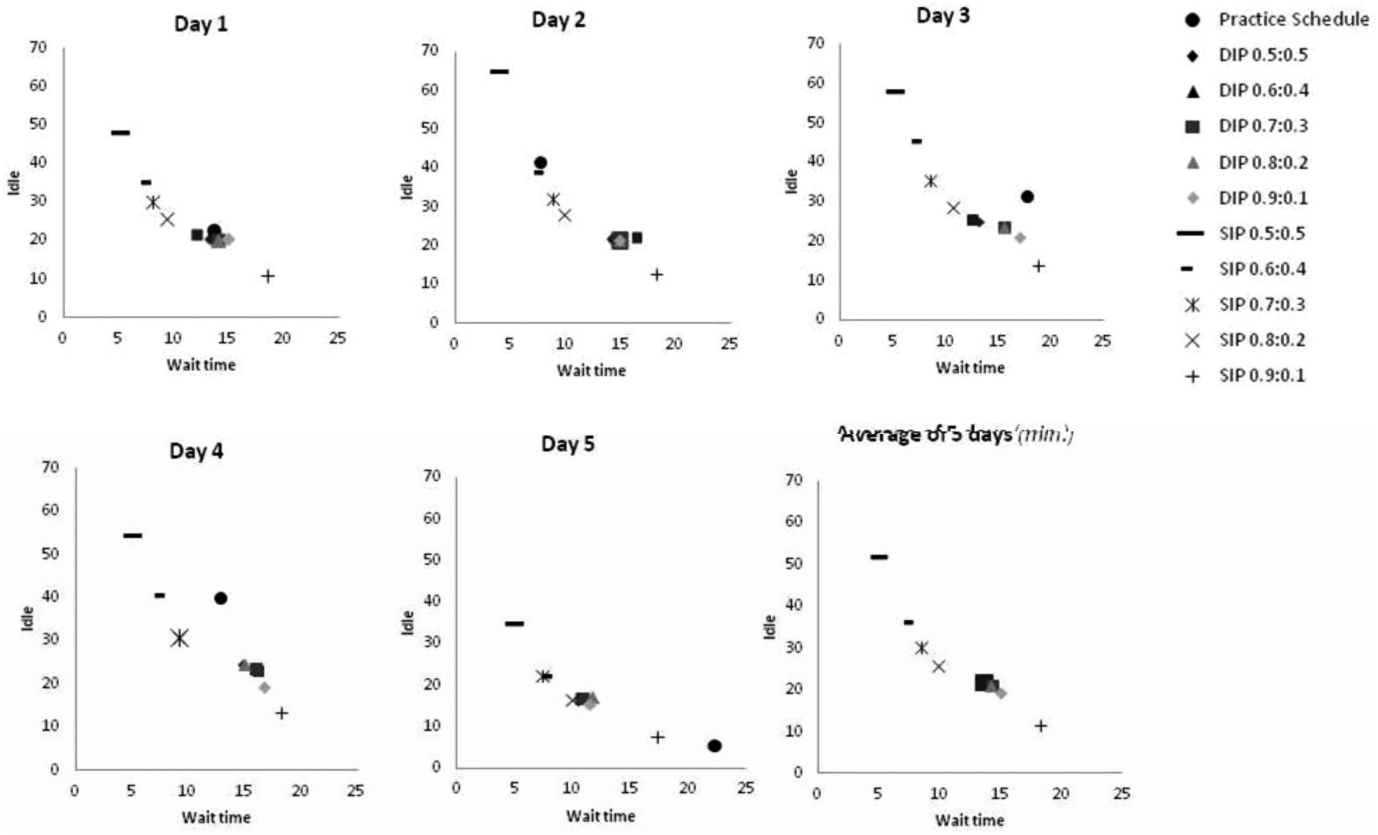

Fig. 6. Expected performance of the schedules using different weight combinations (min.).

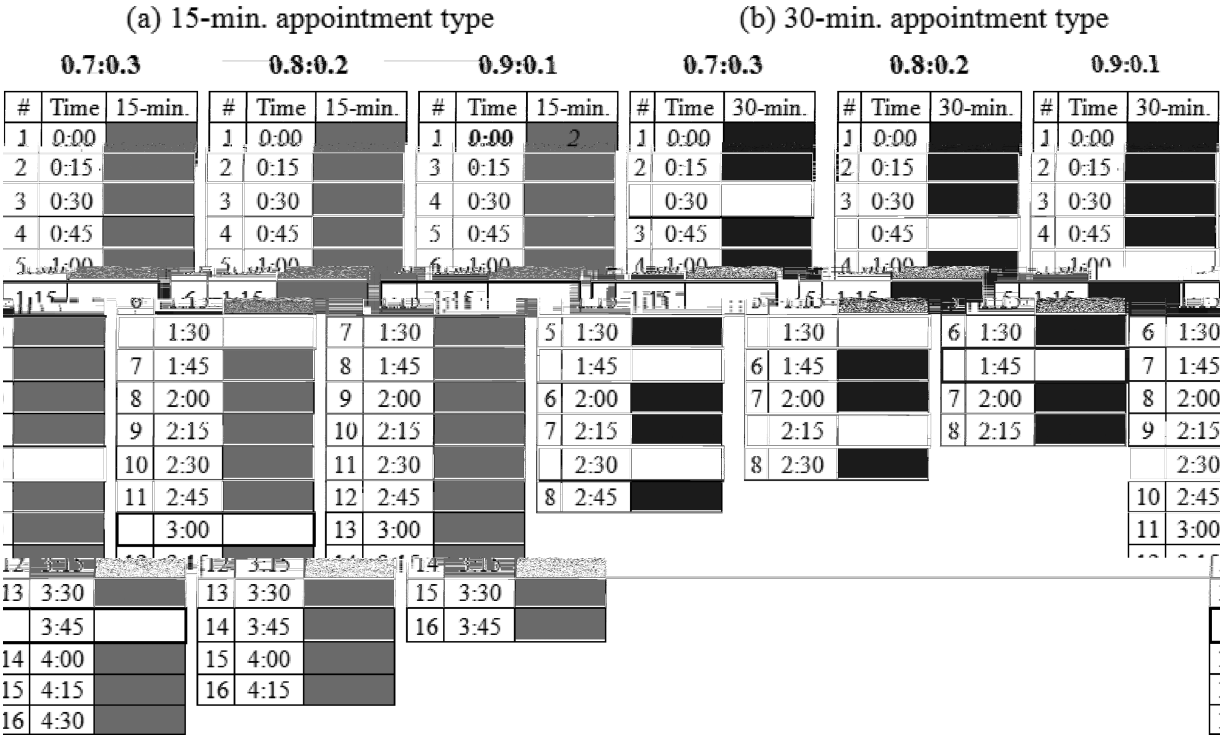

Fig. 7. Optimal SIP schedules for homogeneous patients under different weight combinations. #: Patient number.

5. Computational results

Our computational results consist of five distinct parts. To develop intuition, we first consider the structure of optimal schedules when all patients are of the same type. Second, we look at optimal sequences and appointment times under the DIP and SIP, when a mix of patient types has to be scheduled in a clinic session. In practice, however, sequences need to be flexible so as to accommodate patient preferences and keep the practice financially viable. Hence, in the third part, we look at a variety of heuristic sequences that a practice might prefer, and how slack should be optimally introduced into these sequences to prevent the accumulation of wait time. In the fourth part, we conduct sensitivity on the length of appointment slots and its impact on a practice's performance. Finally, we compare schedules based solely on uncertain provider service time durations—a common practice in the appointment scheduling literature—to our model where both provider and nurse steps, with uncertain service times at both steps, are modeled.

5.1. *Spacing appointment times for homogeneous patients*

We start with the simplest case: How should appointment times be spaced throughout the session if all patients are homogeneous? We consider optimal appointment spacing for three homogeneous patient scenarios for a clinic session: (i) 8 high-complexity (HC) patients; (ii) 16 low-complexity (LC) patients; and (iii) 16 same-day (SD) patients. The optimal schedules for these three scenarios are shown below:

Figure 8 shows that slack is necessary in schedules with HC and LC appointments, but not necessary when all are SD appointments. In the HC case, slack appears after two successive appointments, except at the beginning of the day, where it appears after three successive appointments. Figure 8 also illustrates that LC and SD are indeed different patient categories as we hypothesized: the former needs slack at regular intervals while the latter can do without slack. This is because SD appointments involve less variability in service times with provider (the bottleneck resource) than LC and HC appointments, as shown in Table 3 in section 3.4. As a result, scheduling consecutive SD appointments without slack does not lead to any significant accumulation of patient waiting.

Figure 9 shows the 50th and 90th percentiles of wait time by the patient number in the sequence. As the day progresses, wait time accumulates, but the introduction of slack brings it back down; hence the serrated shapes in the graphs for HC and LC cases. In the HC case, the wait time drops after the third, fifth, and seventh patients, due to slack. In the SD case, there is no slack, so we only have a gradual accumulation of wait time. Note that this accumulation is not as significant compared to the other two cases.

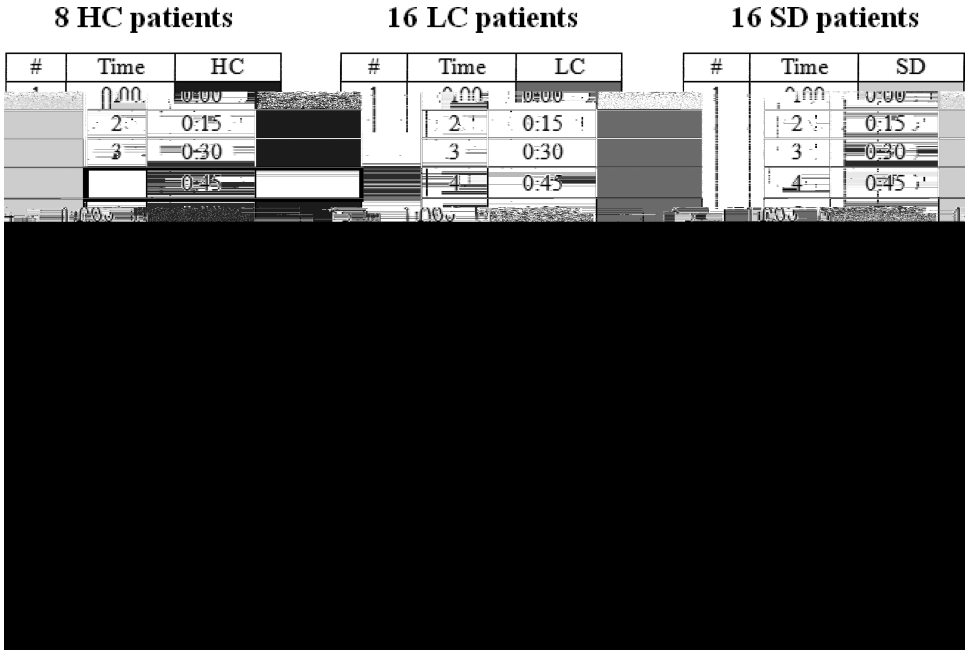

Fig. 8. Spacing appointment times for homogeneous patients.

5.2. *DIP vs. SIP*

Consider Figure 10 (a), which shows the practice schedule for one of the five afternoon sessions observed. In total, 10 patients were scheduled for a provider: three HC patients; three LC patients; and four SD patients. Notice that there is slack after every HC patient. This is in fact the current scheduling policy: the practice uses two 15-min. slots in the calendar for every HC patient. The HC patient scheduled at 2 pm has until 2:30 pm; the HC scheduled at 2:45 pm has until 3:15 and so on.

The afternoon schedules created by the DIP and the SIP models (Figure 10 (b) and (c)) show that there is no need to book slack after every single HC appointment. We see that slack is typically scheduled after two successive HC appointments, consistent with what we found in the homogeneous HC patient case (see previous subsection). In the DIP, HC and LC appointments are double-booked at the 2:15 pm. slot. The empty slot immediately after, at 2:30 pm, provides the necessary time for the provider to see the second patient.

The schedules we observe for the DIP and the SIP models consistently follow the features we see in the example shown in Figure 10. The DIP seems to be *dome-shaped* since it always schedules slack in the middle of the section which implies that the appointment interval lengths increase toward the middle and then decrease to the end of section. This slack in the middle section helps relieve the congestion that naturally accumulates over time. The sequence of the DIP locates HC appointments (with the longest average

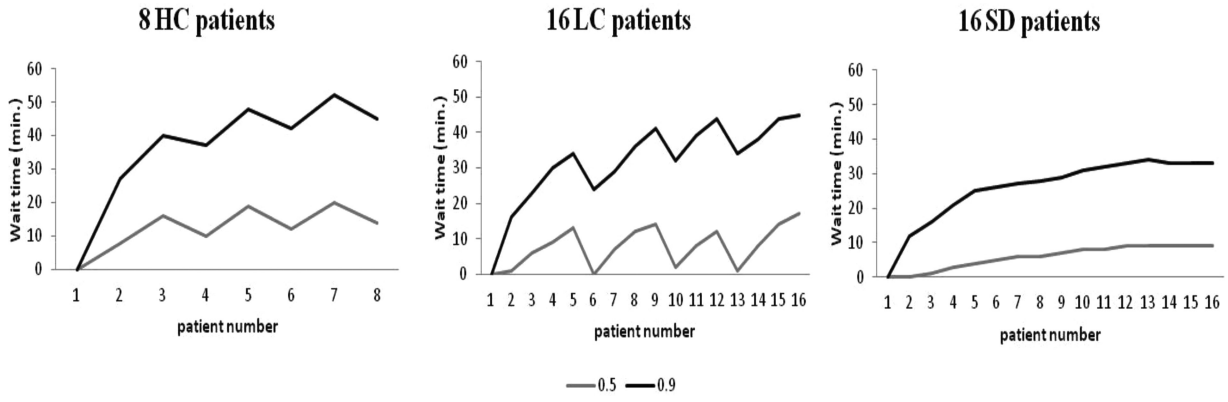

Fig. 9. 50th and 90th percentiles of the patient wait times (min.).

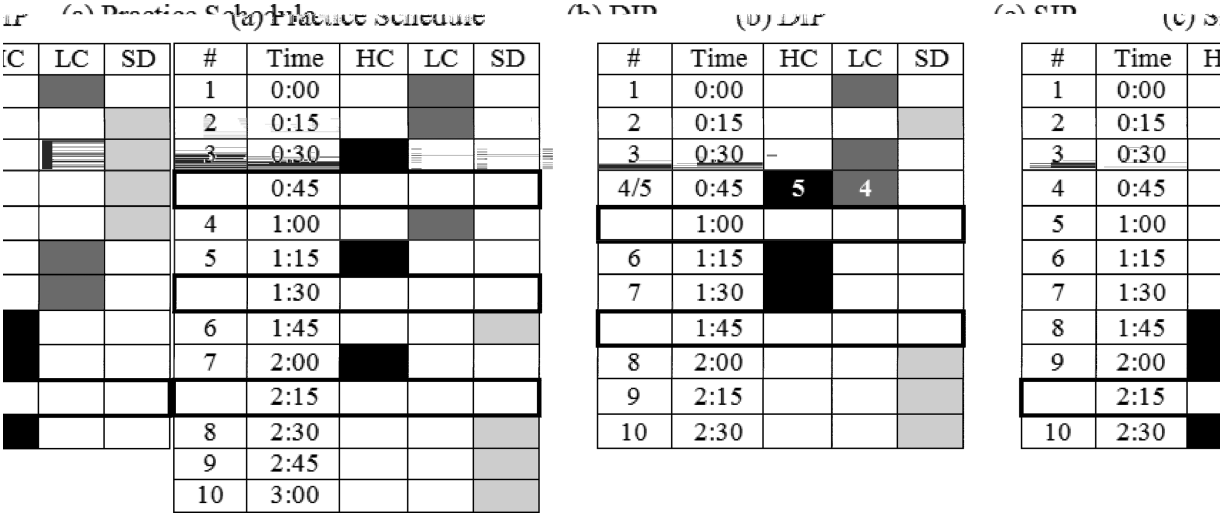

Fig. 10. Schedules associated with one afternoon.

service time) towards the middle, LC towards the beginning, and most SDs towards the end. The SIP, meanwhile, follows, for the most part, the well known *SPT* (shortest processing time) rule. SPT translates to scheduling shortest mean appointments earlier in the schedule. The longer mean appointments, HC appointments, are scheduled towards the end, and the LC appointments are mostly clustered in the middle of the session.

In addition, these SPT sequences are fairly consistent with the sequences generated by the SIP under other different weight combinations that we discuss in Section 4.2. As the idle time weight increases, we observe less slack and more double booking. Also, none of the optimal SIP schedules under different weights start with HC appointments.

Table 4 shows percentage increase in the weighted sum of idle time and wait time. When averaged for five afternoon sessions, the practice's schedule is 24% worse in the objective value compared to the SIP and 16% worse than the DIP. The Value of the Stochastic Solution (VSS), the difference of performance between the SIP and the DIP, is 10%. In terms of total wait times (lobby + exam room), the SIP is 25% better than the practice schedule when averaged over the five afternoon sessions. Furthermore, the 90th percentile of waiting time in the exam room is 20% less in the SIP compared to the practice schedule. To further illustrate this point, Figure 11 displays the 90th percentiles of wait time by the patient number in the different schedules (Practice, DIP, and SIP) for one of the five afternoon sessions.

Table 4. Percent increase in the weighted sum of provider idle time and patient wait time (objective)

Average of 5 days	Practice Schedule vs. DIP	Practice Schedule vs. SIP	VSS
Objective	16%	24%	10%

While the wait time observed by the different patients in the sequence following the practice schedule is highly variable, wait times in the DIP and the SIP increase relatively smoothly. Wait time of the SIP is significantly below that of both the DIP and the practice schedule.

5.3. *Heuristic sequences*

In the two previous subsections, we have identified the structure of optimal schedules for homogeneous sets of patients, as well as a mix of patient appointment types. However, rigid adherence to sequences shown in Figure 10 (b) and (c)—based on the DIP and SIP—are not practical in reality. A dome-shape or SPT sequence is likely to be near optimal, but patients have time of day preferences; it is unrealistic to expect that all patients will be amenable to accepting slots only at a certain time of the day.

To be truly patient centered, therefore, we need to test schedules that provide sufficient flexibility for patients to have time-of-day options. For example, rather than have all HC appointments at the end or the middle of the day,

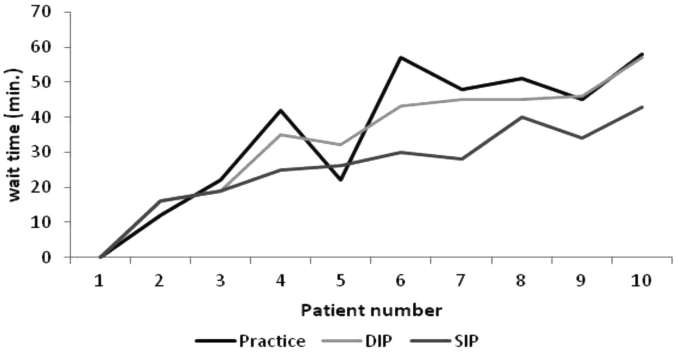

Fig. 11. 90th percentiles of the patient wait times under three different schedules: Practice, DIP, and SIP.

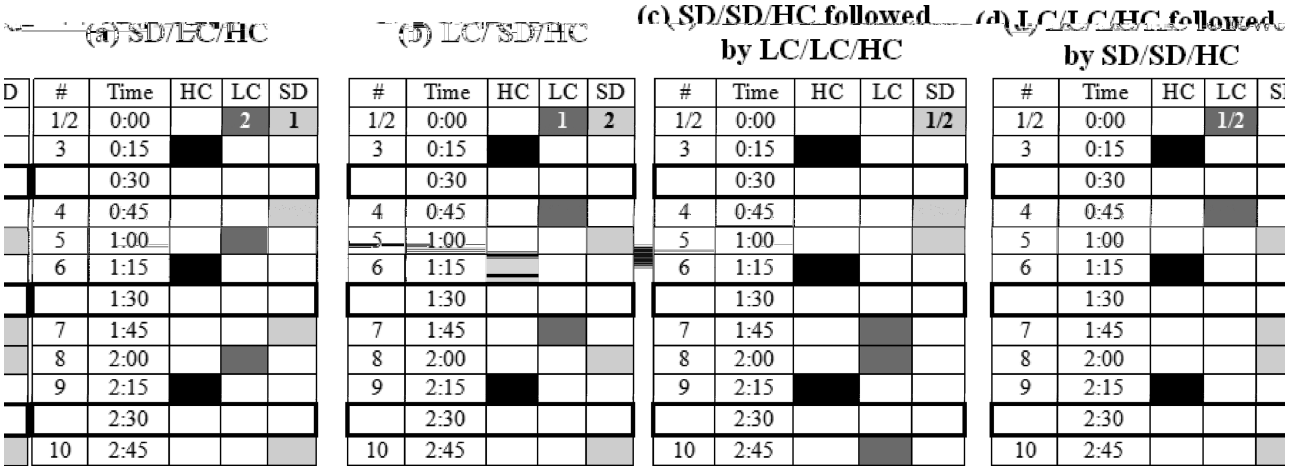

Fig. 12. 3-Appointments-per-Hour (3AH) schedules given optimal appointment times.

the practice may like to make one HC appointment or LC appointment available each hour in a session.

On the other hand, the practice also has to stay financially viable. To do so, each provider in the practice we worked with needed to see at least 3 patients per hour. Hence, to satisfy the practical needs of patients and providers, we now explore a number of sequences that satisfy the 3- Appointments per Hour (3AH) criterion and provide a flexible schedule that allows an option for each type of patient class during every hour of the session. These sequences are shown in Figure 12.

In all four sequences, we have three appointments per hour; we call a block of three appointments a *triad*. The triads are described by the sequence of patient types scheduled. For instance, an SD/LC/HC triad schedules a same-day patient, followed by a low-complexity prescheduled patient and then a high-complexity prescheduled patient. The last appointment of the triad in all four sequences is always a HC appointment. We also examined triads where an HC appointment comes first, but the performance is almost 50% worse than that of the triads where the HC comes last. Hence, we focus on sequences in which a HC appointment is always the last appointment in each triad.

The optimal appointment times for these sequences, which determine the positioning of the slack in the schedule, are obtained using the SIP. In all four sequences, the very first triad in the session involves a double booked slot. This follows Bailey-Welch rule (Bailey, 1952; Welch, 1964). We also see that the SIP consistently suggests the introduction of slack—an empty 15-min. slot—at the end of each triad, in each of the four sequences. The consistency of this pattern is a key finding: if the practice chooses any of the above triad structures for a session, then it is clear where slack should be located.

If we compare the performance of these four schedules (Figure 13) with the SIP in which both the sequence and appointment times are simultaneously optimized (see previous subsection, we find that they are between 9–11% worse in the objective value, when averaged over five afternoon sessions. This may be interpreted as the price of allowing greater flexibility in sequences to accommodate patient preferences. We note, however, that the four heuristic sequences are still 17% better on average than the ad-hoc schedules that were used in the practice.

In addition, we compare the performance of the 3AH schedules shown above with that of optimal schedules under different weights discussed in Section 4.2. We find that the average performance of the 3AH schedules over five sessions is not dominated by the SIP optimal schedules for weight combinations 0.5:0.5, 0.6:0.4 in terms of the two criteria, expecting waiting time and idle time. Indeed, idle time of the 3AH schedules is on average 27 minutes lower than that of the SIP 0.5:0.5 schedules while wait time is only 7 minutes higher.

5.4. *Granularity of appointment slots*

Thus far, we have assumed that our appointment slots are 15 minutes long. Patient appointments will always be at the four quarters of the hour, and therefore easier to remember. But what if the practice tried appointment slots that were 5 minutes long? Patients could be given appointments in 5-min. intervals and allocated a number of consecutive 5-min. slots depending on their needs. The results of such a change would be no worse, since the current 15-min. slot schedules are a feasible solution when the day is broken into 5-min. slots; in fact the schedule might use session time more efficiently. The only inconvenience would be that patients may find appointment times at, say, 9:35 am or 10:55 am, harder to recall and keep track of. We found making slot length more granular does improve the objective value, but only around 4%. The returns do not appear to be significant to justify a change.

What if the minimum slot length was 20 minutes instead of 15 minutes? This means that we are implicitly incorporating greater slack within each appointment. The

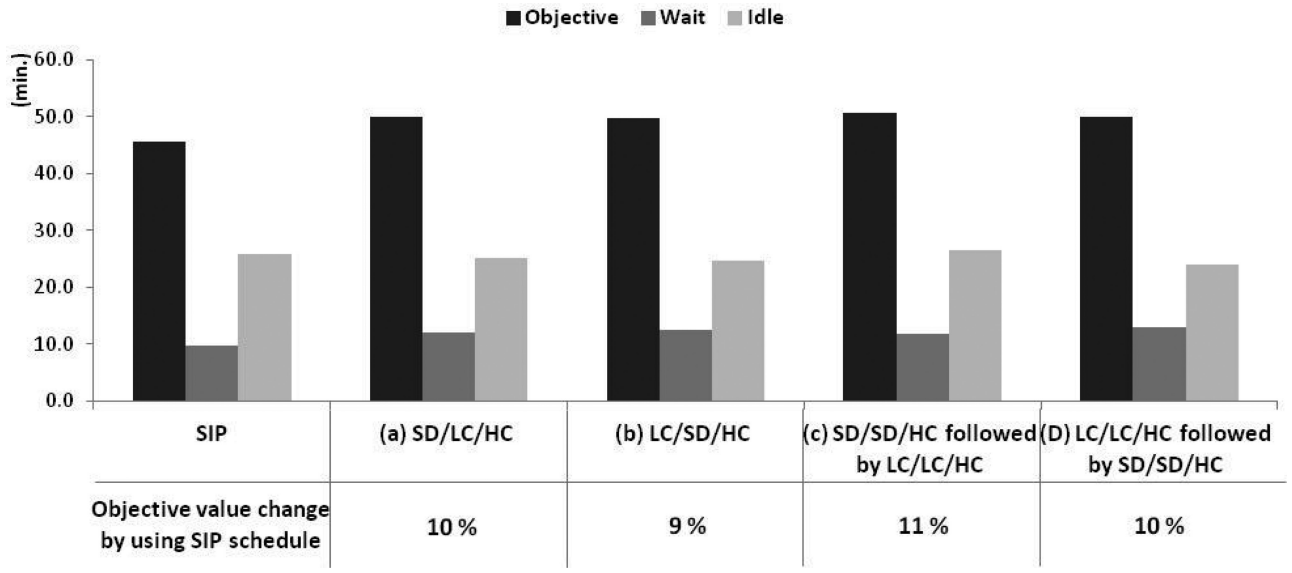

Fig. 13. Performance comparison of the SIP schedule vs. the four 3AH schedules.

performance of such a schedule is 6% worse compared to using 15-min. slots. As shown in Figure 14, we compared the different appointment slot lengths on the five afternoon sessions observed in practice. As appointment slot lengths become more granular, the objective values generated by the SIP are slightly reduced.

5.5. *Comparison with provider-only models*

We compare the performance of two models: (i) the nurse and provider model and (ii) the provider only model. In the integer programs (DIP and SIP), thus, we use both steps, the nurse and provider steps, for (i), but we only account for the provider in (ii). We compare the resulting schedules for each of the 5 afternoon sessions. The average results are summarized in Figure 15.

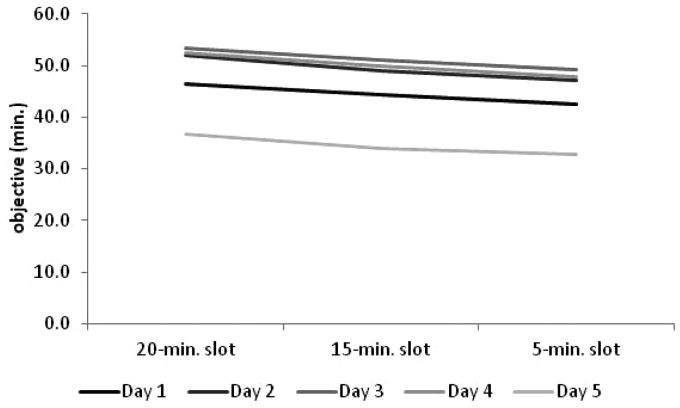

Fig. 14. Weighted idle time and wait time (objective) by the SIP under different appointment slot lengths.

Figure 15 shows that considering the nurse step to generate the optimal schedule results on average in a 21% decrease in the weighted sum of provider idle time and patient wait time. In particular, wait times decrease by 64% which is fairly significant.

Figure 16 shows the schedules generated by the SIP for (i) the nurse and provider model, versus (ii) the provider model. The schedules are significantly different. The provider only schedule starts with a HC appointment at the beginning of the session and includes no slack, resulting in significantly increased patient wait times. Therefore, the nurse step is a critical factor in capturing patient wait times and needs to be considered in outpatient appointment scheduling.

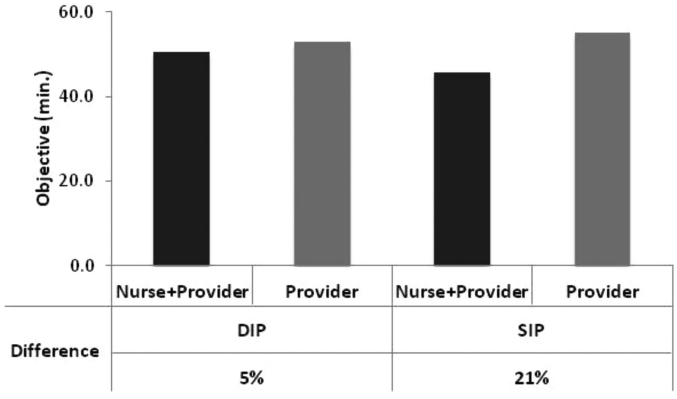

Fig. 15. Performance improvement of nurse and provider steps vs. provider step using DIP and SIP. *Nurse+Provider: the model using service time with both Nurse and Provider; Provider: the model using service time with only Provider.

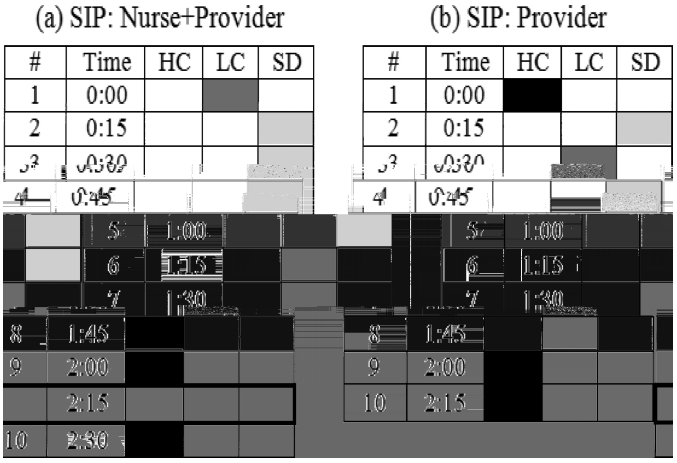

Fig. 16. Schedule for nurse and provider steps vs. provider step using SIP.

6. Conclusion

In summary, our empirical study sheds light on the scheduling challenges facing family care practices. We first use the study to propose a new patient classification scheme. Next, we formulate a stochastic program to model the appointment sequencing and scheduling problem under the new classification and two service steps (nurse and provider). The objective is to minimize a weighted combination of a patient's wait time and a provider's idle time. The model sequences patient types with different nurse and provider time requirements and staggers their appointment times appropriately while keeping the basic slot structure traditionally used by the schedulers at the practice.

The contributions of our research are as follows. First, our new patient classification scheme is meaningful and broadly applicable in primary care. The service time distributions with nurse and provider for specific patient conditions are useful in and of themselves. From an operational point of view, we demonstrate that different amounts of slack are necessary in the schedule depending on the type of patient. It is known that patients with chronic conditions need longer appointments with their providers. Our model provides sufficient space in the schedule for such patients, yet ensures that provider idle time is not more than necessary.

Second, from a modeling point of view, we develop, unlike previous studies, a stochastic program that captures both the patient classification and the entire patient flow through the practice including initial wait, nurse check-up, wait in exam room and provider check-up. Third, we determine the optimal placement of slack (unscheduled slot times) to mitigate the effect of variability of service time with the nurse and the provider, under various patient sequences; this includes sequences that are attractive to the practice because they facilitate the accommodation of patient preferences and yet are financially viable. Our analysis of these sequences shows that optimal appointment times consistently follow a specific structure: an empty slot after every group of three scheduled appointments that includes a 30-min. patient. This results in easy to implement guidelines. Finally, we compare the proposed optimal and heuristic schedules with schedules actually used in practice.

While we were principally interested in the structure of sequences and appointment times, the stochastic program can also be used in a *dynamic* sense. This is important because schedules are not constructed all at once but as calls come in, one at a time. As a companion to the models presented in this paper, we have developed a practical Excel tool that allows the practice to explore the performance of different schedules in real time as patients call in. A Gantt chart in the spreadsheet allows the scheduler to visualize how the appointments are staggered. Double booking of slots is allowed in the tool. The scheduler can dynamically insert new patients into the schedule and obtain the expected performance based on 500 scenarios randomly sampled from the empirical data. The measures provided are: wait time (total as well as by patient position in the sequence), idle time, and finish time. We provide the capability for measuring averages, 50th and 90th percentiles for the current partial schedule. A preliminary version of the spreadsheet is available at: [http://blogs.umass.edu/hyunjuno/]

Our study has some limitations, which provide opportunities for future work. The nine work days chosen for the time-study may not be entirely representative of the volume and mix of patients served. Our main goal, though, was to capture the distribution and variability of nurse and provider service times for different patient types. Comprehensive self-reported data on patient conditions and provider service times does exist in the National Ambulatory Medical Care Surveys (NAMCS) conducted each year. A natural next step would be to check whether the insights of this study apply to such a nationally representative data set.

We do not consider aspects such as no-shows and patient punctuality. But note that no-shows and the probability that a same-day slot goes idle can be modeled by allowing service times with a provider and a nurse to be 0. With some probability, which can be set equal to the no-show rate, these 0-length durations will be randomly picked in the sample average approximation method, in generating the scenarios, and will thereby impact the optimal schedules and performance measures. The assumption that patients arrive punctually can be relaxed by defining a new variable, $\tau_{i,s}^{EL}$, to capture the Earliness/Lateness of patient i under scenario s, and adding it in constraint (6) in the IP model so that the start time of patient i with a nurse is equal to or greater than the earliness/lateness plus appointment time of patient i.

Finally, in this article we consider only schedules for a single nurse and provider. We are however currently working on extending our models and the Excel tool to practices with shared resources. For instance, we consider two nurses

that can flexibly attend to the needs of the patients of two providers.

Acknowledgments

This work was funded in part by from the National Science Foundation (NSF CMMI 1031550). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Bailey, N. (1952) A Study of queues and appointment systems in hospital outpatient departments with special reference to waiting times. *Journal of the Royal Statistical Society*, **14**, 185–199.

Balasubramanian, H., Biehl, S., Dai, L., and Muriel, A. (2103) Dynamic scheduling of same-day requests in multi-physician primary care practices in the presence of prescheduled appointments, online at *Health Care Management Science*, DOI: 10.1007/s10729-013-9242-2

Berg, B. P. (2012) *Optimal Planning and Scheduling in Outpatient Procedure Centers*. [Raleigh, North Carolina], North Carolina State University. http://www.lib.ncsu.edu/resolver/1840.16/7926.

Bodenheimer, T., and Pham, H. H. (2010) Primary care: current problems and proposed solutions. *Health Affairs (Project Hope)*, **29**, 799–805.

Cayirli, T., and Veral, E. (2003) Outpatient scheduling in health care: A review of literature. *Production and Operations Management: an International Journal of the Production and Operations Management Society*, **12**, 519.

Cayirli, T., Veral, E., and Rosen, H. (2006) Designing appointment scheduling systems for ambulatory care services. *Health Care Management Science*, **9**, 47–58.

Cayirli, T., Veral, E., and Rosen, H. (2008) Assessment of patient classification in appointment system design. *Production and Operations Management*, **17**, 338–353.

Chakraborty, S., Muthuraman, K., and Lawley, M. (2012) Sequential clinical scheduling with patient no-show: The impact of pre-defined slot structures. *Socio-Economic Planning Sciences*, **47**(3), 205–219.

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

Chew, S. F. (2011). Outpatient appointment scheduling with variable interappointment times. *Modelling & Simulation in Engineering*, **2011**.

Denton, B., and Gupta, D. (2003) A Sequential Bounding Approach for Optimal Appointment Scheduling. *IIE Transactions*, **35**, 1003–1016.

Denton, B., Viapiano, J., and Vogl, A. (2007) Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health Care Management Science*, **10**, 13–24.

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

Gul, S., Fowler, J. W., Denton, B. T., and Huschka, T. (2011) Bi-criteria scheduling of surgical services for an outpatient procedure center. *Production and Operations Management*, **20**, 406–417.

Hammersley, J. M. and Handscomb, D. C. (1964) *Monte Carlo Methods*. Methuen, London.

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

Ho, C. and Lau, H. (1992) Minimizing Total Cost in Scheduling Outpatient Appointments, *Management Science*, **38**, 1750–1764.

Holland, J. H. (1975) *Adaptation in Natural and Artificial Systems*. The University of Michigan Press, Ann Arbor, MI.

Kaandorp, G., and Koole, G. (2007) Optimal outpatient appointment scheduling. *Health Care Management Science*, **10**, 217–229.

Klassen, K. J., and Rohleder, T. R. (1996) Scheduling Outpatient Appointments in a Dynamic Environment, *Journal of Operations Management*, **14**, 83–101.

Klassen, K. J., and Yoogalingam, R. (2009). Improving performance in outpatient appointment services with a simulation optimization approach. *Production and Operations Management,* **18**, 447–458.

Kleywegt, A., Shapiro, A., and Homem de Mello, T. (2002) The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, **12**, 479–502.

Lin, J., Muthuraman, K., and Lawley, M. (2011) Optimal and approximate algorithms for sequential clinical scheduling with no-shows. *IIE Transactions on Healthcare Systems Engineering*, **1**, 20–36.

Mancilla, C. and Storer, R. (2012). A sample average approximation approach to stochastic appointment sequencing and scheduling. *IIE Transactions*, **44**, 655–670.

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

Robinson, L. W. and Chen, R. R. (2003) Scheduling doctors' appointments: optimal and empirically-based heuristic policies. *IIE Transactions*, **35**, 295–307.

Robinson, L. W. and Chen, R. R. (2011) Estimating the implied value of the customer's waiting time. *Manufacturing Service Oper. Management.* **13**:1, 53–57.

Rockafellar, R. T. and Wets, R. J.-B. (1991) Scenarios and policy aggregation in optimization under uncertainty. *Mathematics of Operations Research*, **16**, 119–147.

Saremi, A., Jula, P., Elmekkawy, T., and Wang, G. (2013) Appointment scheduling of outpatient surgical services in a multistage operating room department. *International Journal of Production Economics*, **141**, 646–658.

Soriano, A. (1966) Comparison of two scheduling systems. *Operations Research*, **14**, 388–397.

Turkcan, A., Zeng, B., Muthuraman, K., and Lawley, M. (2011) Sequential clinical scheduling with service criteria. *European Journal of Operational Research*, **214**, 780–795.

Van Slyke, R. M. and Wets, R. J.-B. (1969) L-shaped linear programs with applications to optimal control and stochastic rogramming. *SIAM Journal of Applied Mathematics*, **17**, 638–663.

Welch, J. D. (1964) Appointment systems in hospital outpatient departments. *Operational Research Quarterly*, **15**, 224–232.