Guidelines for Scheduling in Primary Care under Different Patient Types and Stochastic Nurse and Provider Service Times: An Empirically Driven Mathematical Programming Approach

Oh, Hyun Jung  
Ph.D student, Mechanical and Industrial Engineering, University of Massachusetts, Amherst

Muriel, Ana  
Associate professor, Mechanical and Industrial Engineering, University of Massachusetts, Amherst

Balasubramanian, Hari  
Assistant professor, Mechanical and Industrial Engineering, University of Massachusetts, Amherst

Ptaszkiewicz, Thomas  
Office Manager, Atkinson Family Practice, Amherst, MA

Atkinson, Katherine J.  
MD, Atkinson Family Practice, Amherst, MA

Abstract

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 workday 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 face 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 workday 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 face 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 paper 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 face 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.

Our model is motivated and calibrated using empirical data collected at a 3-provider family medicine practice in Massachusetts. The data was collected using a time-motion study conducted on 9 workdays 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.

**Table 1 Appointment Mix**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of scheduled patients</td>
<td>420</td>
</tr>
<tr>
<td>Total number of observed patients</td>
<td>364</td>
</tr>
<tr>
<td>No-shows</td>
<td>3%</td>
</tr>
<tr>
<td>Cancellations and Reschedules</td>
<td>5%</td>
</tr>
<tr>
<td>Prescheduled appointments</td>
<td>75%</td>
</tr>
<tr>
<td>Same-day appointments</td>
<td>25%</td>
</tr>
<tr>
<td>30-min. appointments</td>
<td>33%</td>
</tr>
<tr>
<td>15-min. appointments</td>
<td>67%</td>
</tr>
</tbody>
</table>

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 U.S. 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 5 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 paper 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) classifies 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 et al. 2011]. To maintain consistency with the literature, we use the broader term outpatient practice instead of primary care practice, and service time instead of face time 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 significantly reduces wait time.

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.

The papers discussed above assume patients to be homogenous. However, the outpatient practice generally consists of various patient types, each of whose face 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. 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 set 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.

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).
We next turn to papers that use stochastic linear programming. Outpatient surgical scheduling is relevant to our work since procedure durations – like face 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 et al. (2011) 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 et al. (2011) 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 procedure 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 to 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 3-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: 1) dome-shaped according to the mean service time ($\mu$), 2) dome-shaped according to adjusted service times ($\mu+\alpha\sigma$), 3) increasing variance, and 4) increasing coefficient of variability of service time.

Muthuraman and Lawley (2008), Chakraborty, Muthuraman and Lawley (2010), Lin, Muthuraman and Lawley (2011), Turkcan et al. (2011), and Chakraborty, Muthuraman, and Lawley (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.

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 or face 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 step 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 paper. 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 face time with nurse and provider for different patient types and ailments.

3.1. Data Collection Methodology

We collected data at a 3-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 workdays: July 7th, 18th, 22nd, August 3rd, 8th, and Oct. 5th, 7th, 8th, 9th 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. So our data is not merely a sample of what happened during a workday; it can be used to reconstruct a complete chronology of patient flow on the 9 workdays.
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: 1) wait time in the lobby; 2) time with a nurse; 3) wait time in the exam room; 4) time with a provider; and 5) 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 face 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, 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 face time with nurses and providers. As shown in Figure 3, however, face time with a medical team (nurses and providers) is highly variable; furthermore, the face 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 Face Time by Patient Conditions

We found that face 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 face time with nurses and providers. Notice that while face time with providers is typically higher, nurse times are non-trivial and in some cases higher. For instance, patients scheduled for WCC 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.

![Box Plots of Face Time with Nurses and Providers by Patient Conditions](image)

* TN: time with nurses  TP: time with providers

**Figure 4 Box Plots of Face Time with Nurses and Providers by Patient Conditions**

3.4. Improved Appointment Classification

The practice currently schedules two types of appointments, 15-min. and 30-min., in 15 minute slots. In the provider’s schedule, a 30-min. appointment takes up 2 consecutive 15 minute slots. The 30-min. appointments consist of routine physical exam; 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.
Table 2 Face time with Nurse and Provider by Patient Type under Current Patient Classification (min.)

<table>
<thead>
<tr>
<th></th>
<th>Nurse</th>
<th>Provider</th>
<th>T-test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>30-min.</td>
<td>19.0</td>
<td>19.0</td>
<td>0.000</td>
</tr>
<tr>
<td>15-min.</td>
<td>9.0</td>
<td>16.0</td>
<td>Ref.</td>
</tr>
</tbody>
</table>

As Table 2 shows, the face times of the patients we observed in our time-study 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 Face time with Nurse and Provider by Patient Type under New Patient Classification (min.)

<table>
<thead>
<tr>
<th></th>
<th>Nurse</th>
<th>Provider</th>
<th>T-test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>HC (High Complexity)</td>
<td>18.0</td>
<td>20.0</td>
<td>0.005</td>
</tr>
<tr>
<td>LC (Low Complexity)</td>
<td>9.0</td>
<td>17.0</td>
<td>Ref.</td>
</tr>
<tr>
<td>SD (Same-day)</td>
<td>10.0</td>
<td>13.0</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Table 3 shows that the differences among the 3 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 3 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-minute 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 workday progresses, whereas all prescheduled patients, whether in the HC or the LC appointments, are known at the beginning of the workday.
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 face 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 1) the sequence in which the patient types are scheduled, and 2) the appointment times of each patient. Because the slots in our case-study are 15-minutes long, the appointment times are always in 15-minute increments. For any feasible first stage decisions (which determine the schedule for the session), nurse and provider face times are realized in the second stage, resulting in idle time for the provider and wait time for the patients.

We create 1000 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 Shapiro 2007]. The two stage stochastic integer program is described formally below:
Notation:

$N$  
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

$A$  
Weight for idle time.

$B$  
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.

$\tau_{n,HC}^{i,s}$  
Face time with a nurse for patient $i$, if of type HC, under scenarios $s$.

$\tau_{n,LC}^{i,s}$  
Face time with a nurse for patient $i$ if of type LC, under scenarios $s$.

$\tau_{n,SD}^{i,s}$  
Face time with a nurse for patient $i$ if of type SD, under scenarios $s$.

$\tau_{p,HC}^{i,s}$  
Face time with a provider for patient $i$ if of type HC, under scenarios $s$.

$\tau_{p,LC}^{i,s}$  
Face time with a provider for patient $i$ if of type LC, under scenarios $s$.

$\tau_{p,SD}^{i,s}$  
Face time with a provider for patient $i$ if of type SD, under scenarios $s$.

Variables

$\tau_{n}^{i,s}$  
Face time of patient $i$ with a nurse under scenarios $s$.

$\tau_{p}^{i,s}$  
Face time of patient $i$ with a provider under scenarios $s$.

$y_{i,s}^{\text{start}}$  
Start time of patient $i$ with a nurse under scenarios $s$.

$y_{i,s}^{\text{finish}}$  
Finish time of patient $i$ with a nurse under scenarios $s$.

$z_{i,s}^{\text{start}}$  
Start time of patient $i$ with a provider under scenarios $s$.

$z_{i,s}^{\text{finish}}$  
Finish time of patient $i$ with a provider under scenarios $s$.

$A_{i} \in \{0,1\}$  
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,\ldots,15 for a 4-hour session.
The problem is modeled as the following integer program.

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

subject to:

\[ x_{0,s}^{\text{finish}} = 0, \quad \forall s \in S \]  

\[ x_{1,s} = 0, \quad \forall s \in S \]  

\[ \tau_{i,s}^{n} = \tau_{i,s}^{n,HC} + \tau_{i,s}^{n,LC} + \tau_{i,s}^{n,SD} + C_{i}, \quad \forall i \in I, s \in S \]  

\[ \tau_{i,s}^{p} = \tau_{i,s}^{p,HC} + \tau_{i,s}^{p,LC} + \tau_{i,s}^{p,SD} + C_{i}, \quad \forall i \in I, s \in S \]  

\[ y_{i,s}^{\text{start}} \geq 15X_{i}, \quad \forall i \in I, s \in S \]  

\[ y_{i,s}^{\text{finish}} \geq y_{i-1,s}^{\text{start}}, \quad \forall i \in I, s \in S \]  

\[ y_{i,s}^{\text{finish}} = y_{i,s}^{\text{start}} + \tau_{i,s}^{\text{nurse}}, \quad \forall i \in I, s \in S \]  

\[ z_{i,s}^{\text{start}} \geq y_{i,s}^{\text{finish}}, \quad \forall i \in I, s \in S \]  

\[ z_{i,s}^{\text{finish}} \geq z_{i-1,s}^{\text{start}}, \quad \forall i \in I, s \in S \]  

\[ z_{i,s}^{\text{finish}} = z_{i,s}^{\text{start}} + \tau_{i,s}^{\text{provider}}, \quad \forall i \in I, s \in S \]  

\[ \sum_{i=1}^{n} A_{i} = N_{HC} \]  

\[ \sum_{i=1}^{n} B_{i} = N_{LC} \]  

\[ \sum_{i=1}^{n} C_{i} = N_{SD} \]  

\[ A_{i} + B_{i} + C_{i} = 1 \]  

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

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 face 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 face 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 time-of-day preferences, it is important to derive robust guidelines on where slack should be strategically positioned in the schedule.

In this paper, we do not consider no-shows for the prescheduled appointments or the fact that a SD slot can go idle because no patients called in for that slot. But note that the model can easily include both these aspects. No-shows and probability that a same-day slot goes idle can be modeled by allowing face times with a provider and a nurse to be 0. With some probability, which can be set equal to the no-show or idle 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.

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 provider idle time had a higher weight than patient waiting. This makes intuitive sense since idle time is experienced by a single person while wait time accumulates across all scheduled patients.

The results are shown in Figure 5. 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 how wait times accumulate as the day progresses. The SIP, which considers variability, exhibits greater sensitivity toward changes in weights. We noticed that the SIP 0.7:0.3 schedules provided low wait times but more provider idle time in the schedule than would be acceptable for the practice, while the SIP 0.9:0.1 schedules provided very little slack and thereby increased patient waiting more than would be acceptable. The 0.8:0.2 SIP schedules provided 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 6 (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 combination in the remainder of our computational experiments.
Figure 6 Optimal SIP schedules for homogeneous patients under different (provider idle time : patient wait) weight combinations

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 face time durations – a common practice in the appointment scheduling literature – to our model where both provider and nurse steps, with uncertain face 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: 1) 8 high-complexity (HC) patients; 2) 16 low-complexity (LC) patients; and 3) 16 same-day (SD) patients. The optimal schedules for these three scenarios are shown below:

![Figure 7 Spacing appointment times for homogeneous patients](image)

Figure 7 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 3 successive appointments. Figure 7 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.
Figure 8 50th and 90th percentiles of the patient wait times over the course of the session (min.)

Figure 8 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 3rd, 5th and 7th patients, due to slack. In the SD case, there is no slack, so we only have a gradual accumulation of wait time. However, this accumulation is not as significant compared to the other two cases.

5.2. DIP vs. SIP

Consider Figure 9 (a), which shows the practice schedule for one of five afternoon sessions observed. In total, 10 patients were scheduled for a provider: 3 HC patients; 3 LC patients; and 4 SD patients. Notice that there is slack after every HC patient. This is in fact the current scheduling policy: the practice uses two 15 minute 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 schedules created by the DIP and the SIP models (Figure 9 (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, we have double booked with HC and LC appointments at a 2:15 pm slot. The empty slot immediately after, at 2:30 pm, provides the necessary time for the provider to see the next patient.
The schedules we observe for the DIP and the SIP models consistently follow the features we see in the example shown in Figure 9. The DIP seems to be *dome-shaped* since it always schedules HC appointments (with the longest average face 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.

![Schedule of one afternoon example](image)

**Figure 9 Schedule of one afternoon example**

<table>
<thead>
<tr>
<th>#</th>
<th>Time</th>
<th>HC</th>
<th>LC</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13:30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>13:45</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>14:00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>14:15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>14:30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>15:00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>15:15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>16:00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>16:15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>16:30</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>#</th>
<th>Time</th>
<th>HC</th>
<th>LC</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13:30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>13:45</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>14:00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>14:15</td>
<td>5</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>14:30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>15:00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>15:15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>15:30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>15:45</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>16:00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Objective</th>
<th>Practice Schedule vs. DIP</th>
<th>Practice Schedule vs. SIP</th>
<th>VSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average of 5 days</td>
<td>16%</td>
<td>24%</td>
<td>10%</td>
</tr>
</tbody>
</table>

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 5 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 10 displays the 90th percentiles of wait time by the patient number in the different schedules (Practice, DIP, and SIP) for one of the 5 afternoon sessions. 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.

![Graph showing patient wait times](image)

**Figure 10** 90th percentiles of the patient wait times over the course of one of the afternoon sessions under three different schedules: Practice, DIP, and SIP

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 9 (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 time-of-day options. For example, rather than have all HC appointments at the end or the middle of the day, the practice may like to make one HC appointment or LC appointment available each hour in a session.

For SD patients, calls have to be fulfilled at a short notice. 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 allows the practice to provide a patient who calls in at, say, 8 am with a late morning slot. Indeed, the practice we worked with has followed this policy based on our recommendation. Figure 11 shows the average number of prescheduled and same-day appointments by time of day for 9 workdays with 2 doctors 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.

![Figure 11 Pre-scheduled vs. Same-day by Time of Day](image)

In addition to addressing patient preferences, 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, we now explore a number of sequences that satisfy the 3 Appointments per Hour (3AH) criterion but also are sensitive to patient preferences. These sequences are shown in Figure 12.

![Figure 12 3-Appointments-per-Hour (3AH) schedules given optimal appointment times](image)

In all four sequences, we have 3 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 location 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 minute slot – at the end of each triad, in each of the 4 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.

![Figure 13 Performance comparison of the SIP schedule vs. the four proposed 3AH schedules](image)

<table>
<thead>
<tr>
<th>Objective</th>
<th>Wait</th>
<th>Idle</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a) SD/LC/HC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(b) LC/SD/HC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(c) SD/SD/HC followed by LC/LC/HC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(d) LC/LC/HC followed by SD/SD/HC</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Objective value change by using SIP schedule</th>
</tr>
</thead>
<tbody>
<tr>
<td>-10%</td>
</tr>
<tr>
<td>-9%</td>
</tr>
<tr>
<td>-11%</td>
</tr>
<tr>
<td>-10%</td>
</tr>
</tbody>
</table>

If we compare the performance of these four schedules with the SIP in which both the sequence and appointment times are simultaneously optimized (see previous subsection, Figure 9 (c)), we find that they are between 9-11% worse in the objective value, when averaged over 5 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.
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 minute intervals and allocated a number of consecutive 5 minute slots depending on their needs. The results of such a change would be no worse, since the current 15 minute slot schedules are a feasible solution when the day is broken into 5 minute 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 performance of such a schedule is 6% worse compared to using 15 minute slots. As shown in Figure 14, we compared the different appointment slot lengths on the 5 afternoon sessions observed in practice. As appointment slot lengths become more granular, the objective values generated by the SIP slightly reduce.

![Figure 14 Weighted provider idle time and patient wait (objective) generated by the SIP under different appointment slot lengths on 5 afternoon sessions](image)

5.5. Comparison with Provider-Only Models

We compare the performance of two models: 1) the nurse and provider model and 2) the provider only model. In the integer programs (DIP and SIP), thus, we use both steps, the nurse and provider steps, for 1),
but we only account for the provider in 2). We compare the resulting schedules for each of the 5 afternoon sessions. The average results are summarized in Figure 15.

* Nurse+Provider: the model using face time with both Nurse and Provider
* Provider: the model using face time with only Provider

**Figure 15 Performance improvement achieved by considering nurse and provider steps vs. only the provider step in optimizing the schedules using both deterministic and stochastic approaches.**

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 Schedules generated by the SIP for (a) the nurse and provider model and (b) the provider only model**
Figure 16 shows the schedules generated by the SIP for (a) the nurse and provider model, versus (b) 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.

6. Conclusion

In summary, our empirical study sheds light on the scheduling challenges facing small family care practices across America. 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 face 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 face 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 3 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/](http://blogs.umass.edu/hyunjuno/)

Finally, we are 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.

**Acknowledgements**

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**


