Stochastic Appointment Scheduling in a Team Primary Care Practice with Two Flexible Nurses and Two Dedicated Providers

Hyun-Jung Alvarez-Oh, Hari Balasubramanian, Ekin Koker, Ana Muriel

1. Introduction

Effective scheduling in primary care practices plays an important role in smoothing patient flow. Many papers have studied the scheduling problem in the outpatient setting but commonly assume a single step in the patient flow process: the provider step. However, most practices also involve a nurse step before the provider step. According to our empirical data analysis in Oh et al. (2013), nurse service time durations for many appointments are comparable to provider service time durations. For example, for routine physical and well child examinations—two common appointment types in primary care—nurses spend as much time with the patients as do providers. In addition, Oh et al. (2013) reports that there is a significant difference in the performance as well as the structure of the optimal schedule in a single-provider practice when the nurse step is explicitly considered in the scheduling formulation compared with when it is not.

Another common assumption is a single resource at each step: for example, a solo provider working at the practice. However, the majority of practices (68%) have two or more providers (Bodenheimer and Pham 2010). In our consultations with practices, we have noticed that nurses work as a team in prepping patients for provider appointments. Nurses flexibly see patients scheduled on providers’ calendars whenever they are available, whereas providers stay dedicated to their appointment schedules. We call this a team primary care practice.

This multistep patient flow process with multiple human resources at each step coupled with uncertain service times makes the problem difficult from an optimal scheduling viewpoint. Figure 1 shows an example of a team primary care practice. Patient waiting can occur in the lobby and the examination (hereafter, “exam”) room; that is, there are two queuing steps in the patient flow process. Each provider has a set of patients whose appointment times determine the sequence in which they are seen by a nurse; but because the nurses can see any of the two providers’ patients, a crossover can happen in the schedule when the nurse’s service time for a particular patient is long. In Figure 1, a longer than expected nurse service time for patient 3 results in patient 5 seeing the other nurse and potentially completing the nurse step and being seen by the provider earlier.
Choosing to keep the original sequence versus following the new crossover sequence has implications for both patient waiting as well as the idle time of the provider. From a queuing viewpoint, this is a choice of queue discipline at the second step of a tandem queue. The queue discipline could be either to see patients in the original appointment sequence (no crossovers allowed) or to see patients on a first come, first served basis (allow crossovers). Although we have observed that situations that lead to crossover are common in practice, it is not clear what their operational impact is or what a practice’s strategy should be.

Our contribution is a new two-stage stochastic integer programming formulation of the team primary care practice that allows for nurse flexibility and patient crossovers while minimizing a weighted combination of provider idle time and patient wait time. To the best of our knowledge, appointment scheduling in team primary care practices has not been tackled from a mathematical programming perspective. Modeling nurse flexibility and patient crossovers is nontrivial when we consider that the appointment times need to be optimally determined in the first stage, and the resulting patient flows through the practice need to be identified. Computationally the problem becomes more challenging given a set of probabilistic nurse and provider service time scenarios that get realized in the second stage. For a feasible first-stage solution of appointment times, sequence changes due to crossover can happen in some scenarios, whereas in other scenarios the original sequence will be retained. Thus, tracking patient flow in each scenario within the framework of a stochastic program poses a modeling challenge, which we tackle in this paper.

The team practice model helps answer the following practically relevant questions:

- Do optimal team practice schedules consistently exhibit a certain structure that translates to generalizable guidelines?
- How does the (flexible) team practice perform in comparison with a practice in which each nurse is dedicated to a single provider?
- What if we impose that, despite the possibility of crossover, the provider sees patients in the same sequence as their original appointment times? How would the wait and idle times of such a solution compare with a solution that did allow crossovers?
- How do patient no-shows and greater variability in service times affect optimal schedules?

From the computational tractability viewpoint, we demonstrate the use of tightening constraints and a lower bounding procedure to solve realistic instances with a large number of scenarios.

The rest of the article is structured as follows. In Section 2, we provide a brief literature review, particularly focusing on studies considering multiple resources and steps, and exact solution approaches. In Section 3, we describe the team practice and introduce the mathematical model and solution method. Our computational study is divided into two parts: in Section 4, we derive practical guidelines relevant for practices, and in Section 5, we discuss computational feasibility. In Section 6, we extend our model to incorporate no-shows. In Section 7, we summarize our conclusions.

2. Literature Review

Outpatient appointment scheduling is a widely studied topic. It refers to a broad class of problems ranging from primary care to specialty clinics and outpatient procedure centers (nuclear medicine and chemotherapy infusion centers) to surgical scheduling. The characteristics of each setting have led to a unique set of assumptions and modeling approaches in the literature. For a comprehensive review of recent literature related to optimization in appointment scheduling, we refer the reader to Ahmadi-Javid et al. (2017).
Papers in appointment scheduling can be broadly split into three categories based on the time horizon modeled: on the more operational side, papers that focus on a single day, optimizing direct wait time of the patients at the clinic; on the more tactical side, papers that focus on multiple days/weeks, optimizing indirect wait time between an appointment request and an actual appointment day; and papers that focus on a combination of both. In the latter two categories, rather than provide a comprehensive review, we will instead provide a few representative examples. We will explore in more detail papers that focus on a single day, which are the most closely related to our study. Furthermore, we restrict ourselves to papers that consider multiple resources and multiple steps in the patient flow process and use an exact solution approach.

Zacharias and Armony (2017) combine tactical- and operational-level scheduling by jointly optimizing panel size (the number of patients a primary care physician or a practice cares for in the long term) and the number of offered appointments per day. Their objective is to minimize clinical delay (direct wait time) and appointment delay (indirect wait time), and by using a single-server, two-stage queueing model (an appointment book and the clinic itself) they demonstrate that an “Open Access” policy, whereby the clinic tries to offer a same-day appointment with high probability, is optimal. Wang and Gupta (2011) consider assigning dynamically arriving phone requests for appointments with patient preferences to multiple primary care providers (i.e., a single step with multiple servers). The problem is motivated as a Markov decision process (MDP) and solved with heuristics. Other papers that consider multiple days or weeks in their model and multiple stages include Pérez et al. (2011, 2013) (nuclear medicine clinics) and Conforti et al. (2010) and Castro and Petrovic (2012) (radiotherapy treatment).

We now discuss papers that schedule a single day. El-Sharo et al. (2015) focus on the overbooking aspect of single-stage, multiserver outpatient scheduling in the presence of no-shows, cancellations, and walk-ins. Wang and Fung (2014a, 2014b) and Wang et al. (2015) study outpatient scheduling with patient preferences using different approaches (MDP, Integer Programming [IP], and Dynamic Programming [DP], respectively). All of their work involves a single-stage queue with multiple servers (doctors). Tsai and Teng (2014) approach the online scheduling problem for physical therapy in a rehabilitation service that involves a single stage but multiple servers, such as therapy equipment. Balasubramanian et al. (2014) consider assigning dynamically arriving same-day requests to multiple primary care providers (i.e., a single step with multiple servers) to maximize the number of same-day patients seen in the day. They use a stochastic dynamic programming approach. Hahn-Goldberg et al. (2014) consider a multistage, multiserver chemotherapy scheduling problem and use the deterministic version to come up with templates that are dynamically adjusted, also considering future requests. Lin (2015) considers an appointment scheduling problem in an eye clinic where the patients go through different pathways (stages) and use different servers (doctors, nurses, optometrists) and propose an adaptive scheduling heuristic with memory (previous distinct schedules are recorded). Liang and Turkcan (2016) focus on a single-stage, multiserver queue in an oncology clinic and propose mixed integer programming models for nurse assignment and patient scheduling. Their unique contribution is that they consider patient acuity levels that estimate nurse requirements more accurately.

Surgical scheduling research differs substantially from the rest of the outpatient appointment research because it can be done in both an inpatient and an outpatient setting. However, the outpatient surgical scheduling closely relates to our study. Three examples for outpatient surgical scheduling with multiple steps and resources are Saremi et al. (2013), Bai et al. (2016), and Neyshabouri and Berg (2017). Saremi et al. (2013) consider three stages of the operating room: preoperation, surgery, and recovery in the postanesthesia care unit (PACU). Bai et al. (2016) tackle the multiple-OR and PACU surgery scheduling problem by using a sample-gradient based algorithm. Neyshabouri and Berg (2017) consider two stages of the operating room: surgery and surgical intensive care unit (SICU). They decompose the two stages into separate mixed integer linear programs and use a column and constraint generation algorithm. SICU length-of-stay (LOS) is in a different time-scale from PACU stay and surgery duration because SICU LOS might take a few days, whereas PACU stay and surgery duration usually take only hours. These two time-scale features make their problem different from those in Saremi et al. (2013) and Bai et al. (2016). For a more extensive review of recent research on outpatient surgical scheduling with multiple steps and resources, we point to the literature review section in Neyshabouri and Berg (2017). Our paper differs from surgical scheduling because the shared (flexible) resources are upstream in primary care (nurses), as opposed to downstream in surgery (PACU or SICU beds). Patient crossovers (first come, first served [FCFS] in second stage) are a direct consequence of this fact and make the problem more challenging.

The two papers, in addition to Oh et al. (2013), that relate most closely to our work are Castaign et al. (2016) and Kuiper and Mandjes (2015). Castaing et al. (2016) formulate the outpatient scheduling problem in chemotherapy infusion centers as a two-stage stochastic program. One interesting aspect of the problem is that the infusion chairs are considered as the main resource and the nurses are considered as an external resource; nurses can care for multiple patients at the same time. To reduce waste of expensive chemotherapy drugs, a common practice is to delay the preparation of a dose until the patient is ready. The objective is to minimize a weighted combination of
wait time and total length of operations, which directly correlates with staff overtime. Because of the weakness in their formulation due to big-M type constraints, they face high run times for the solution procedure and propose decomposition heuristics that perform well.

Kuiper and Mandjes (2015) approach the outpatient appointment scheduling problem as a tandem-type queuing model with two stages and single server at each stage. The objective is to minimize a weighted combination of patient wait time and provider idle time. They approximate the service times with their phase-type counterparts, which fit the distribution using first two moments. They propose a recursive method to evaluate the sojourn-time distribution of patients and computationally determine optimal schedules in a transient as well as steady-state environment. They also consider extensions such as heterogeneous service-time distributions and blocking, in which the buffers between two stages are finite (clients cannot move between two stages because the servers are busy). They observe the familiar dome-shaped pattern in their optimal schedules.

Finally, in our previous work in Oh et al. (2013), we formulate a two-stage stochastic integer program for the single-provider primary care practice composed of a single nurse and provider. The model is optimally scheduled and sequenced patient appointments using stochastic service time of two service steps, nurse and provider, and new patient classifications with the objective of minimization of patient wait time and provider idle time. We suggest the scheduling guidelines obtained by the optimal schedules as well as heuristic schedules capable of accommodating patient time-of-day preferences.

All of the papers above either consider 1) a single stage with multiple servers, 2) multiple stages with a single server, 3) multiple stages and servers but with a single server at each stage, or 4) multiple stages and servers but deal with the problem deterministically. None of them deal with crossovers. This is understandable because problems involving multiple steps with multiple resources at each step, stochastic service durations, and crossovers are intractable using exact approaches. We tackle this issue by exploiting the structural properties of the problem with two servers and by developing tightening constraints and lower bounding techniques.

In contrast to Oh et al. (2013) and the rest of the literature, this paper contributes to current research on outpatient appointment scheduling by formulating a two-stage stochastic programming model (an exact approach) for a two-stage problem (nurse and provider stages) with two servers at each stage (two flexible nurses and two dedicated providers) in a primary care practice. We also consider a homogeneous mix of patients in this study in contrast to the heterogeneous (different patient classifications) in our previous work (Oh et al. 2013). To the best of our knowledge, the problem of appointment scheduling in primary care practices with multiple stages and servers under stochasticity with flexible resources and patient crossovers (FCFS in second stage) has not been tackled before using an exact approach.

3. Modeling Approach

3.1. Description of Team Primary Care Practice

We consider appointment scheduling for a team primary care practice in which two flexible nurses share patients, whereas each provider oversees appointments only from his/her own panel (the configuration shown in Figure 1). At the practice that motivated this study, there are morning and afternoon sessions distinguished by a lunch break. Each session consists of appointment slots, and the length of each slot is 15 minutes. The patient visit consists of the following steps: after check-in, a patient waits in the lobby until a nurse calls (wait time in the lobby); the first available nurse calls the patient into the exam room and examines the patient (nurse service time); after the nurse step, the patient waits in the exam room until her/his primary provider is available (wait time in the exam room); and once the provider finishes with the previous patient, she/he will examine the patient (service time with provider). A provider takes care of the earliest available patient from his own panel after the nurse step. In other words, the provider sees patients in the order of their finish times at the nurse step (FCFS), instead of the order of appointment times. Patient crossover will thus occur when a patient with an earlier appointment may have a long nurse service and end up seeing the provider after a patient with a later appointment.

For service time distributions, we use two sources: empirical and lognormal. Figure 2 shows empirical distributions of service time with nurse and provider for high complexity (HC) patient visits. These data were collected at the collaborating practice as part of the time study in Oh et al. (2013). Type HC involves physicals and complex conditions, which require long service time with nurses and providers. Community health centers based in low-income and medically underserved neighborhoods often schedule a full day of HC appointments for primary care providers.

As shown in Figure 2, service times with both nurse and provider are highly variable. On average, a type HC appointment takes 17.8 minutes with nurse (standard deviation: 10.7 minutes; coefficient of variation $[CV] = 0.6$ minutes) and 19.5 minutes with provider (standard deviation: 8.2 minutes; $CV = 0.42$ minutes). Although provider service times tend to be longer, the service time distribution for the nurse step is skewed to the right, leading to nurse visits that are significantly longer than the provider visits. It is apparent that the nurse and provider steps should be effectively coordinated to avoid long patient waits or low provider utilization.
In addition to random samples of empirical data on HC appointments, we also use samples from lognormal distributions for nurse and provider service time distributions. The lognormal distribution allows us to control for the mean while increasing the variance of service times, thereby allowing us to test a wider range of cases and establish the generalizability of our findings.

3.2. Integer Programming Formulation
We formulate a mixed integer program to schedule patients into appointment slots. Key features of the model are to accommodate two sequential steps (nurse and provider), multiple human resources at each step, stochastic service times, flexible nurses, providers dedicated to own panels, and patient crossover. We use a fixed, predetermined appointment length of 15 minutes and consider homogeneous patients. The objective of the model is to minimize a weighted measure of provider idle time and patient wait time across all scenarios. We assume that the patients punctually arrive at the appointed time because 89% of patients come early or on time according to our data analysis.

We will use the following notation to formulate the problem.

Sets
- $I$: Set of patients to be scheduled in the session, indexed by $i = 1, \ldots, I$
- $J_k$: Set of patients to be scheduled with provider $k$, indexed by $j = 1, \ldots, J_k$
- $S$: Set of scenarios, indexed by $s = 1, \ldots, S$
- $K$: Set of providers, indexed by $k = 1, 2$

Parameters
- $\alpha$: Weight for idle time
- $\beta$: Weight for wait time
- $\tau_{i,s}^n$: Service time of patient $i$ with nurse under scenario $s$
- $\tau_{j,s}^k$: Service time of the $j$th patient to see provider $k$ under scenario $s$
- $f[j,k]$: Patient index (in the overall set of patients in the practice) of the $j$th patient of provider $k$

Variables
- $y_{i,s}^{\text{start}}$: Start time of patient $i$ with nurse under scenario $s$
- $y_{i,s}^{\text{finish}}$: Finish time of patient $i$ with nurse under scenario $s$
- $t_{j,s}^k$: Finish time with nurse of the $j$th patient in provider $k$’s panel under scenario $s$
- $z_{i,s}^{\text{start}}$: Start time of the $j$th patient to visit with provider $k$ under scenario $s$
- $z_{j,s}^{\text{finish}}$: Finish time of the $j$th patient to visit with provider $k$ under scenario $s$
- $N_{i,s}^{\text{max}}$: Maximum of the finish times of patients $1, \ldots, i$ with nurses under scenario $s$
- $P_{j,s}^{\text{max}}$: Maximum of the finish times of patients $1, \ldots, j$ of provider $k$’s panel under scenario $s$
- $X_i$: Appointment slot assigned to patient $i$, an integer variable in $\{0,1,2,\ldots\}$.

The team practice problem, which we refer to as Problem TP, is modeled as the following integer program.

Min. \[ \frac{1}{S} \left[ \alpha \left( \sum_{k} \sum_{s} \left( \sum_{i=1}^{I_k} \frac{1}{\tau_{i,s}^n} - \sum_{i=1}^{I_k} \frac{1}{\tau_{i,s}^n} \right) \right) \right] + \beta \sum_{s} \sum_{i=1}^{N_{i,s}^{\text{max}}} \left( y_{i,s}^{\text{start}} - 15X_i \right) + \sum_{s} \left( \sum_{k} \sum_{j=1}^{J_k} \left( z_{j,s}^{\text{start}} - t_{j,s}^k \right) \right) \]
For minimum constraints (5, 8, 12, and 15), we apply a big-M method and introduce two sets of binary variables:

\[
\begin{align*}
\text{Subject to} \\
y_{s,i}^{\text{start}} &= 0 \quad \forall s \in S, i = 1, 2 \\
z_{k,s}^{\text{finish}} &= 0 \quad \forall k \in K, s \in S \\
X_i &= 0 \quad i = 1, 2 \\
y_{s,i}^{\text{start}} &\geq \min(y_{s,1}^{\text{finish}}, y_{s,2}^{\text{finish}}) \quad \forall s \in S \\
N_{i,s}^{\text{max}} &\geq \max(y_{1,s}^{\text{finish}}, y_{2,s}^{\text{finish}}) \quad \forall s \in S \\
N_{i,s}^{\text{max}} &\geq \max(N_{i-1,s}^{\text{max}}, y_{i-1,s}^{\text{finish}}) \quad \forall i \in 4..I, s \in S \\
y_{s,i}^{\text{finish}} &\geq \min(N_{i-1,s}^{\text{max}}, y_{i-1,s}^{\text{finish}}) \quad \forall i \in 4..I, s \in S \\
y_{s,i}^{\text{finish}} &= y_{s,i}^{\text{start}} + z_{i,s}^{N} \quad \forall i \in I, s \in S \\
y_{s,i}^{\text{start}} &\geq 15X_i \quad \forall i \in I, s \in S \\
t_{k,s}^{f} &= y_{f_{s}}^{\text{finish}} \quad \forall k \in K, j \in J_k, s \in S \\
z_{k,s}^{\text{start}} &\geq \min(t_{k,s}^{f}, t_{2,s}^{f}) \quad \forall k \in K, s \in S \\
p_{k,s}^{\text{max}} &\geq \max(t_{k,s}^{f}, t_{2,s}^{f}) \quad \forall k \in K, s \in S \\
p_{j,s}^{\text{max}} &\geq \max(p_{j,s}^{\text{max}}, p_{j+1,s}^{\text{max}}) \quad \forall k \in K, j \in (3..J_k), s \in S \\
z_{j,s}^{\text{start}} &\geq \min(p_{j,s}^{\text{max}}, p_{j+1,s}^{\text{max}}) \quad \forall k \in K, j \in 2..J_k - 1, s \in S \\
z_{k,s}^{\text{start}} &\geq p_{k,s}^{\text{max}} \quad \forall k \in K, s \in S \\
\sum_{j \in J_k} z_{j,s}^{\text{finish}} &= z_{j,s}^{\text{start}} + t_{j,s}^{f} \quad \forall k \in K, j \in J_k, s \in S \\
z_{j,s}^{\text{finish}} &\geq z_{j-1,s}^{\text{finish}} \quad \forall k \in K, j \in J_k, s \in S \\
X &\geq 0, \text{ INT}; \quad y_{s,i}^{\text{start}}, y_{s,i}^{\text{finish}}, z_{i,s}^{\text{start}}, z_{i,s}^{\text{finish}} \geq 0.
\end{align*}
\]

The objective function (1) minimizes a weighted average measure of provider idle time and patient wait time across all scenarios. Note that provider idle time is calculated as the finish time of the last patient minus the sum of the service times of all patients with provider \( k \) under each scenario. The wait time in the lobby is the difference between the patient’s start time with nurse and the appointment time. The wait time in the exam room is calculated as the sum of the differences of the patients’ start times with provider and finish times at the nurse step. Constraints (2–4) initialize the start time with nurses for the first two patients, and set the 0th patient finish time with provider \( k \) to be zero in every scenario. Constraint (5) makes sure that provider 3 is seen by the earliest available nurse, by comparing the finish times of the first two patients with nurses. Constraint (6) calculates the maximum finish time of the first two patients with nurses. Similarly, Constraint (7) keeps track of maximum finish time with nurse for patients 1 to patient \( i - 1 \). The maximum value for patients 1 through \( i - 2 \) is compared with the finish time of patient \( i - 1 \) with nurses in Constraint (8) to find the earliest time a nurse is available to take care of the subsequent patient \( i \). Constraint (9) calculates the finish time of patient \( i \) with nurse, as the start time plus the service time with nurse. Constraint (10) ensures that a nurse can only see a patient after the patients’ appointment time (recall that patients arrive punctually; they are not available any earlier or later than their appointment time). Constraint (11) assigns the nurse finish time of patient \( i \) in the full schedule to the finish time with nurse of the corresponding patient \( j \) in provider \( k \)’s panel for each scenario \( s \); that is, it transfers information from the full ordered set of patients in the practice to the ordered set of patients in doctor \( k \)’s panel. Constraints (13) and (14) track the maximum of the nurse finish times of the first \( j - 1 \) patients scheduled from provider \( k \)’s panel, and this maximum value is recursively updated in Constraint (15). Constraints (12) and (15) ensure that each provider \( k \) serves the patient \( j \) who finishes the nurse step earlier; this is done by comparing the nurse finish times of the first \( j + 1 \) patients in provider \( k \)’s panel, to account for possible crossover. Constraint (16) ensures the start time of the last patient seen by provider \( k \) is no sooner than the finish time with nurse for all the patients in the panel. Constraint (17) calculates the finish time of patient \( j \) as the start time plus service time with provider \( k \). Constraint (18) ensures that provider \( k \) starts to examine the \( j \)th patient after seeing the \( j - 1 \)th patient.

The current model with minimum and maximum constraints can be reformulated into a linear program as follows. The maximum constraints can be easily broken into two inequalities, one for each term in the maximum. For minimum constraints (5, 8, 12, and 15), we apply a big-M method and introduce two sets of binary variables:
otherwise.

Each of the minimum constraints is reformulated into two constraints. Let $M^1$ and $M^2$ be sufficiently large constants. Constraint (8) is rewritten as Constraints (8-1) and (8-2):

$$y^\text{start}_{i,s} \geq N_{i-1,s}^\text{max} - M^1 n_{i,s} \quad \forall i \in \{4..N\}, s \in S$$

$$y^\text{finish}_{i,s} \geq y^\text{start}_{i-1,s} - M^1 (1 - n_{i,s}) \quad \forall i \in \{4..N\}, s \in S$$

Similarly, Constraint (15) becomes Constraints (15-1) and (15-2):

$$z^k_{i,s} \geq p^k_{j-1,s} - M^2 p^k_{j,s} \quad \forall k \in K, j \in \{2..J_k - 1\}, s \in S$$

$$z^k_{i,s} \geq z^k_{i+1,s} - M^2 (1 - p^k_{j,s}) \quad \forall k \in K, j \in \{2..J_k - 1\}, s \in S$$

Constraints (5) and (12) follow the same structure.

### 3.3. Tightening of the Formulation

The proposed integer programming model is computationally challenging because the number of scenarios needs to be sufficiently high to ensure robustness of the solution. We use 1,000 scenarios in our experiments because it provides a good balance between robustness and computational complexity.

For instances with more than five patients per provider, the general model fails to find a guaranteed optimal solution within four hours of computation time. We thus seek strategies to tighten the formulation. Specifically, we derive tight lower bounds on the big-M parameters and propose a lower bound on the optimal cost based on solving the problem for exhaustive and mutually exclusive subsets of scenarios, stage-based bounds, and additional constraints, to eliminate unnecessary processing and strengthen the formulation. As we shall see in the computational section, this significantly helps reduce the computational time.

First, we tighten the big-M constraints, Constraints (5) and (8) with $M^1$ and Constraints (12) and (15) with $M^2$. $M^1$ is a bound on the difference of nurse finish times of patient $i + 1$ and the maximum of patients $1$ through $i$; and $M^2$ is a bound on the difference of nurse finish times between provider $k$'s patient $j$ and $j + 1$. The following theorems provide closed-form expressions for tight values of $M^1$ and $M^2$, respectively. The proofs are provided in the e-companion.

**Theorem 1.** The value of $M^1$ for each patient under each scenario can be given by

$$M^1_{i,s} = \max\{\tau_{i-1,s}^N + \max(0, 30 - \tau_{i-2,s}^N), \max_{r=1,\ldots,i-2}\{\tau_{r,s}^N - \sum_{u=r+1}^{i-1} \tau_{u,s}^N\}\}$$

**Theorem 2.** The value of $M^2$ for patient $j$ of provider $k$ under scenario $s$ can be provided by

$$M^{2k}_{j,s} = \max\{\tau_{i+1,s}^N + \tau_{i,s}^N + \max(0, - \tau_{i,s}^N + 30), \max_{r=1,\ldots,i}\{\tau_{r,s}^N - \sum_{u=r+1}^{i+1} \tau_{u,s}^N\}\}$$

where $i = f[j,k]$; that is, $i$ is the patient number in the overall practice schedule corresponding to the $j$th patient in provider $k$’s schedule.

Next, we derive a tight lower bound on the optimal solution as follows. Let $S$ be the full set of scenarios, that is, $S = \{1,2,\ldots,N\}$.

- **Step 1:** Divide the set of scenarios $S$ into a number of exhaustive and mutually exclusive subsets $\{S_1, S_2, \ldots, S_n\}$ such that $S = S_1 \cup S_2 \cup \ldots \cup S_n$.
- **Step 2:** Solve Problem TP under each scenario subset $S_i$. Let $C^*_i$ denote the optimal cost, and let $S_i$ be the size of set $S_i$, $i = 1, 2, \ldots, n$.
- **Step 3:** Calculate the lower bound on Problem TP under the full set of scenarios $S$ as $C^*_S = \frac{1}{n} \sum_{i=1}^{n} S_i C^*_i$.

In addition, we propose stage-based lower bounds (see Santos et al. 1995) for both the nurse and provider steps. At the nurse step we derive lower bounds for the start time and finish time with nurses for each patient under each scenario $s$. At the provider step, we determine bounds on the finish time of the last patient with each provider $k$. 

\[ ni_1 = \text{if the earliest nurse available to see patient } i \text{ is the one that serves patient } i - 1, \text{ that is, there is some earlier patient that is still seeing the other nurse; } 0, \text{ otherwise;} \]

\[ p_{i,k}^j = \text{if crossover occurs, that is, the } j\text{th patient to see provider } k \text{ is the } j + 1 \text{ patient in his appointment schedule; } 0, \text{ otherwise.} \]
which is essentially session completion time of each provider. Our lower bounds are derived using Constraints (5)–(9) to calculate the start time and finish time with nurses without consideration of the appointment times introduced in Constraint (10). In other words, the earliest time a patient visit starts can be calculated recursively as the second largest value of the finish times up to patient \( i - 1 \) at the nurse step. This provides tight lower bounds for the nurse start and finish times of patient \( i \) in the nurse step and the completion time with provider \( k \) in the provider step under scenario \( s \).

We also introduce additional constraints to further tighten the formulation and reduce unnecessary processing. First, the appointment times can be required to be in ascending order, without loss of generality (w.l.o.g.); that is, the appointment time of patient \( i + 1 \) must be greater than or equal to that of patient \( i \).

\[
X_i \leq X_{i+1} \leq \cdots \leq X_I, \quad \forall i \in I
\]  

(19)

Second, when appropriate, we restrict the appointment schedule to have at most one open slot (slack) between consecutive patients, both within the overall set of patients in the practice (Constraint (20)) and within the patients in a provider \( k \)'s panel (Constraint (21)), with \( i \) and \( i + 2 \) as consecutive patients in provider \( k \)'s panel. Observe that Constraint (21) does not allow for double booking within provider \( k \)'s panel. This is appropriate for the complex appointments, type HC, under consideration, because they require long service times; double-booking would highly increase patient wait time. Note also that we are assuming that all patients show up at their appointment time. When no-shows are prevalent, this constraint will be relaxed to allow for double-booking.

\[
\begin{align*}
X_{i+1} - X_i & \leq 2, \quad \forall i \in I \\
1 \leq X_{i+2} - X_i & \leq 2, \quad \forall i \in I
\end{align*}
\]  

(20)  

(21)

We must note that the initial model, without these additional constraints, is fully general and yields solutions that follow these properties. These assumptions are the result of observing the properties of the optimal schedules for small instances. The associated constraints are then added to solve the larger instances. They are also backed up by our analysis of the time data obtained from the observed practice. According to our data, 12% of the patients take more than 30 minutes (two slots) and only 3% more than 45 minutes with nurse. Ten percent of the patients spend more than 30 minutes with provider, and the maximum service time for the provider is 44 minutes.

4. Scheduling Guidelines

In this section, we use the optimal schedules to derive guidelines that practices can follow to better balance patient wait and provider idle time. In particular, we identify when to add slack, the quality of schedules that ignore the stochastic nature of nurse and provider visit times, and the impact on performance of sharing nurses across providers and allowing for crossovers.

In our experiments, we use both empirical and lognormal visit time distributions and study small, medium, and large instances: 5 patients, 8 patients, and 10 patients per provider, respectively. These values are chosen on the basis of observed practice, average service times of 20 minutes, and the typical four-hour length of morning and afternoon sessions. Considering the initial time with the nurse, the slack between patients, and the variability of service times, the large instances are reasonably sized. The small instances are chosen as half the size of the large instances, and medium instances are chosen as a value in between. Please note that although the observed practice is unlikely to have only five patients per provider per session, because of computational challenges the small instances are analyzed because it allows us to find the exact optimal solutions. We also include results for medium instances to show the progression of the shape of optimal policies as the number of patients increases. Numerical results, however, are only presented for either small instances, because the results are optimal, or practical large instances that show how the effect of different factors becomes more pronounced.

In the objective function, we use coefficients of 0.8 for idle time and 0.2 for wait time (cost ratio = 4) because we find these weights align best with the desired performance of the practice in our study. In Section 4.6, we explore the sensitivity of the results to this cost ratio.

4.1. Scheduling Guidelines for HC Appointments

Figure 3 below shows the optimal schedules for small (5 patients per provider), medium (8 patients per provider), and large (10 patients per provider) instances with HC appointments following the empirical nurse and provider
visit times observed in practice. The first important structural property of these schedules is that they are staggered (i.e., the slack for each of the two providers is not scheduled for the same appointment interval).

The times given above indicate when the patients are asked to arrive to the practice relative to the practice’s working hours. For example, if the practice opens up at 9 a.m., the patient scheduled at time 0:00 will arrive at 9 a.m. In the practice we have observed, providers are busy with paperwork and other necessary tasks before their first and after their last patients. It is only the idle time between patients that causes inefficiency.

In our effort to derive scheduling guidelines, we compare three schedules: practice policy schedule, identical schedule, and optimal staggered schedule, according to our observation above. The practice policy schedule follows the scheduling rules of the practice that inspired our study. Their policy is to book an HC appointment in two 15-minute slots, because they regard HC appointments as 30-minute appointments; in other words, a 15-minute slack is placed after every HC appointment. The identical schedule is determined by the solution of our model with an additional Constraint (22), which makes sure that both providers have identical schedules.

$$X_i = X_{i+1}, \quad \forall i \in 1, 3, 5$$

Figure 4 displays the schedules of practice, identical, and staggered policies for small instances.

As shown in Figure 4, the identical schedule consists of three appointments followed by slack and two appointments. The first three appointments are consecutively scheduled because the wait time has not accumulated yet. In the staggered schedule, the schedule of provider 1 follows the identical schedule, whereas the schedule of provider 2 assigns slack after two appointments; staggering in this fashion allows a steadier flow into the flexible nurse step. For the team practice under study, our model suggests to schedule two HC appointments followed by
slack, a similar scheduling structure we have proposed in Oh et al. (2013) for the single-provider practice. In addition, because our patient indexing makes the $j$th patient of provider 1 be patient $i = 2j - 1$ and that of provider 2 be patient $i = 2j$ in the overall practice sequence, and thus gives nurse priority to the patients of provider 1 over provider 2 given the same appointment times, the schedule for provider 1 is more packed (i.e., has fewer and later empty slots).

Next we compare the wait time, idle time, and completion time performance of the three schedules: practice policy, identical, and staggered. The values specified below are given for small instances, but medium and large instances also follow similar results. The identical schedule provides approximately 25% better objective value and 45% lower idle time compared with the practice policy, on average over the 1,000 scenarios. In the practice policy, however, the wait time performance is significantly better (3.5 minutes per patient), whereas the average idle time is more than one hour with only five patients per provider. The practice schedule introduces more than enough slack, which causes very low wait times but unsustainably high idle times. The objective difference between the identical and staggered schedules, however, is only 2%. This is because although the staggered schedule improves 17% on wait time, the idle time increases 5% compared with the identical schedule. The numerical values for each policy are given in Table 1.

To provide a better perspective, we display the performance of the practice as the session unfolds, for each of the 10 patients in the sequence. Figures 5 and 6 show the wait time per patient and idle time between patients for all three schedules. In the figures, we omit the first patient for each provider, patients p1 and p2 in the practice, because they are always scheduled at the beginning of the session and independent of the appointment policy used.

Figure 5 shows that the wait time per patient followed by the practice policy is way below the 15-minute line but providers go idle more than 10 minutes before seeing each patient, on average. It is because unneeded slack is scheduled, which results in inefficient use of resources. In the identical and staggered schedules, the wait time accumulates and then drops down where slack has been added. The patient wait time of the staggered schedule stays consistently around the 15-minute line. Thus, patients in the staggered schedule experience less wait time than those in the identical schedule: three patients wait slightly more than 15 minutes in the staggered schedule, whereas five patients wait more than 15 minutes in the identical schedule.

The idle time of both the identical and the staggered schedules (Figure 6) is in a similar range and much less than 10 minutes per patient after the very first two patients. Next we study the 90th percentile of wait time per patient and idle time between patients for the three schedules, to see how they perform in the “worst case.”

Figures 7 and 8 show that each patient’s wait time in the practice policy’s worst case is approximately 30 minutes below the wait times associated with the identical and staggered policies. On the other hand, the provider idle time in the practice policy’s worst case is almost twice that of the identical and staggered. Comparing wait time between the identical and the staggered policies, only two patients in the staggered schedule wait more than 45 minutes,

### Table 1. Comparison of Results for Practice, Identical, and Staggered Policies—Small Instance (Five Patients per Provider) and Empirical Service Times

<table>
<thead>
<tr>
<th></th>
<th>Practice policy</th>
<th>Identical</th>
<th>Staggered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wait time (per patient)</td>
<td>3.45</td>
<td>13.03</td>
<td>10.79</td>
</tr>
<tr>
<td>Idle time (per provider)</td>
<td>62.38</td>
<td>34.09</td>
<td>35.76</td>
</tr>
<tr>
<td>Objective function value</td>
<td>106.72</td>
<td>80.59</td>
<td>78.81</td>
</tr>
</tbody>
</table>

Note. Time values expressed in minutes.
whereas four patients spend more than 45 minutes to wait in the identical schedule. Therefore, the staggered schedule performs fairly well.

In summary, we derive the following guidelines for the scheduling of HC appointments in the practice under study: (1) team practices are better off staggering slack slots rather than locating them identically in both providers’ schedules; (2) the patients of the provider with nurse priority will have a more packed schedule (less slack); (3) two HC appointments should be followed by a slack slot, except perhaps in the first sequence of the session; and (4) no double-booking for a provider in the absence of no-shows, because HC appointments have long and highly variable service times.

4.2. Effect of Service Time Distribution and Variance

To generalize our analysis and insights, we test instances with lognormally distributed nurse and provider service times. This allows us to assess the effect of the distributional shape while keeping the service time mean and variance constant, as well as the effect of increasing the service time variance while keeping the mean the same.

Figure 9 below shows the optimal schedules for small (5 patients per provider), medium (8 patients per provider), and large (10 patients per provider) instances with lognormally distributed service times, with the same mean and variance as in the empirical distribution.

As can be seen above, the structures of optimal schedules in instances with lognormally distributed service times are very similar to those with empirically distributed service times, and the guidelines we developed for HC appointments still hold.

To test the effect of service time variance on the structure of optimal schedules, we created two more sets of instances with lognormally distributed nurse and provider service times with doubled variance (CV = 0.59 for provider and 0.85 for nurse) and quadrupled variance (CV = 0.84 for provider and 1.2 for nurse), respectively. The mean is kept constant. As the service time variance increases, the optimal schedule gets more packed (i.e., less slack is introduced); see Figure 10. Intuitively, this makes sense as the practice is placing a higher weight (0.8) on idle time relative to wait time (0.2).

4.3. Value of Stochastic Solution

The stochastic nature of nurse and provider visit times results in a computationally challenging, large-scale problem including a wide set of scenarios. What would be the loss in performance associated with the schedule generated by a deterministic model that simply considers average nurse and provider visit times? Value of stochastic solution (VSS) is calculated as the difference between the expected cost (over all scenarios) of the schedule generated by the deterministic, or single scenario, version of the problem with expected nurse and provider times, and the cost of the optimal schedule suggested by the stochastic problem. VSS is under 2% of the objective function value of the stochastic solution for all cases, which shows that the deterministic solution is actually a very good
heuristic for this problem. Interestingly, the schedules generated by the deterministic model are also staggered, which further demonstrates the superiority of staggered schedules and explains the high quality of the deterministic solution. The deterministic model thus provides a very effective and efficient heuristic for the case of interest to our practice, a cost ratio of provider idle time to patient wait time of 4, where packed schedules are attractive. The VSS, however, increases as more weight is placed on patient wait, being approximately 10% when the ratio is 1 (0.5:0.5).

4.4. Effect of Nurse Flexibility

Next, we compare the joint performance in wait time and idle time between two independent single nurse–provider practices and a two-flexible-nurse, two-provider team practice, seeing five patients per provider. Table 2 displays the objective function values, mean and 90th percentile of wait time per patient, and idle time per provider for the two practices under empirical and lognormal service time distributions; that is, wait time is averaged across all patients, and idle time is averaged across the providers. Then, the mean and 90th percentiles are found across 1,000 scenarios. The results given from this point on will focus on staggered policy as the optimal solution. We must note that we compared the effect of nurse flexibility statistically using a paired t-test for the means of the objective function over 1,000 scenarios and found that the difference is significant ($p$-value $\approx 0$) under all settings.

Table 2 shows that although the wait time and idle time performance of the team practice (flexible nurses) dominates that of the single-provider practices (a.k.a. dedicated nurses), the impact is rather low from an operational viewpoint. The objective function improvements are 6.5% and 7.5% for empirical and lognormal service time distributions, respectively. However, the benefits of nurse flexibility increase with the increase in the number of patients and the variance in service time, as illustrated by Tables 3 and 4. Table 3 shows the 5 patients per provider and 10 patients per provider cases under lognormally distributed service times. Although the objective function improvements are similar in percentage, 7.5% and 6.5% for 5 and 10 patients per provider, respectively, the 90th percentile improvements in wait time and idle time are more pronounced, with a reduction of 5.5 minutes in
wait time and 6.5 minutes in idle time for the case with 10 patients per provider, whereas it is 6.5 minutes reduction in idle time and 1.5 minutes increase in wait time for the case with 5 patients per provider.

Finally, Table 4 shows the regular and quadrupled service time variance cases for lognormal distribution for 10 patients per provider. The objective function improvement increases from 6.5% for the regular service time variance to 18% for the quadrupled service time variance. Wait time and idle time improvements are also much more pronounced, reaching 8 and 10 minutes for the mean and 21 and 22.5 minutes for the 90th percentile.

4.5. Effect of Crossovers

Crossovers naturally happen in the case of flexible nurses as a later patient in the schedule of a provider may complete the nurse step before an earlier patient with a longer nurse visit time. In our model, we assume that the provider will see next the patient that first becomes available, thus allowing for patient crossover, to minimize uncomfortable patient wait time in the exam room. What would be the loss of performance if crossover is disallowed and providers see patients in the same sequence as they arrive at the practice and are seen by the nurses? Observe that the no-crossover model is easier but not trivial to formulate and solve because tracking patient flow at the flexible nurse step still requires the introduction of binary variables and M constraints; see the e-companion for the detailed formulation. The provider step, however, follows the original patient sequence and is straightforward. As a result, the second set of binary variables and M constraints (in the original team care problem with crossovers) are no longer necessary, and solution speed is improved. In what follows, we compare the joint performance in objective function value, wait time, and idle time. We must note that we compared the effect of crossovers statistically using a paired $t$-test for the means of the objective function over 1,000 scenarios and found that the difference is significant ($p$-value $\approx 0$) under all settings.

Table 5 displays the objective function value, mean, and 90th percentile of wait time and idle time for a team practice with crossover versus a team practice without crossover for five patients per provider for empirically and lognormally distributed service times.

Table 5 shows that although the wait time and idle time performance of the practice with crossovers dominates that of the practice without crossovers, the impact is rather low from an operational viewpoint. The benefits of patient crossover, however, increase with the number of patients and service time variance, as shown in Tables 6 and 7. Table 6 shows the cases with 5 patients per provider and 10 patients per provider under lognormally distributed service times. Although the objective function improvements are similar in percentage, approximately

| Table 2. Comparison of Results for Dedicated vs. Flexible Nurses—Small Instances (Five Patients per Provider) for Empirical and Lognormal Service Times |
|-----------------------------------------------|----------------|----------------|----------------|----------------|
|                                              | Empirical distribution |                  | Lognormal distribution |                  |
|                                              | Dedicated nurses | Flexible nurses | Dedicated nurses | Flexible nurses |
| Mean wait time                              | 11.93           | 10.79           | 11.75           | 10.35           |
| Mean idle time                              | 37.68           | 35.76           | 37.12           | 34.92           |
| 90th percentile of wait time                | 27.41           | 25.21           | 22.01           | 23.52           |
| 90th percentile of idle time                | 63              | 56              | 60              | 53.55           |
| Objective function value                    | 84.15           | 78.81           | 82.89           | 76.57           |

Note. Time values expressed in minutes.
5% for both 5 and 10 patients per provider, the 90th percentile improvement in idle time is significantly more pronounced for the larger-size problem, with savings of almost 10 minutes, whereas the wait time is slightly worse, increasing by approximately one minute relative to the no-crossover solution.

Finally, Table 7 shows the performance comparison for the cases of regular versus quadrupled service time variance with lognormal distribution and 10 patients per provider. The objective function improvement of the schedule with versus without crossovers increases from 5% for the regular service time variance to 14.5% for the quadrupled service time variance. The value of the model with crossovers to reduce both wait time and idle time is significantly higher for quadrupled service time variance. The wait time is reduced by 3 minutes for mean and 4 minutes for 90th percentile, whereas the idle time is reduced by 17 minutes for mean and 24 minutes for 90th percentile.

Although the impact of crossovers is more significant on idle time under the cost ratio of 4 under consideration, further tests using a cost ratio of 1 show even greater improvements on wait time. In the most extreme case, with quadrupled service time variance and 90th percentile, the improvement is 38%, from 33 minutes to 20 minutes. These results can be found in the e-companion to our paper.

In general, the probability of occurrence of patient crossover increases when the number of patients per provider increases and when the variance in the service time increases. In the small instance with empirical distribution, there was only a 5% chance that a particular patient will experience crossover. Therefore, it is not a surprise that a schedule that considered crossover did not have a significant impact. On the other hand, in the large instance with lognormal service time distribution and quadrupled service time variance, there is a 15% chance that a particular patient in the schedule will experience crossover. A patient flow model that accounts for crossovers is therefore more beneficial in the latter situation.

### 4.6. Sensitivity to Cost Ratio and Granularity of Appointment Slots

The e-companion contains further sensitivity analyses with respect to the idle to wait time cost ratio and the appointment slot length. Increasing the weight placed on wait time, by varying the idle to wait time cost ratios from the original (0.8:0.2) to (0.66:0.34) and (0.5:0.5) has the expected results. The number of open slots increases. The original practice schedule, leaving one open slot after every patient, then becomes attractive. Increasing the granularity of the appointment slot lengths by reducing the slot duration from the original 15 minutes to 5 minutes, and even further allowing unrestricted appointment times, results in just modest improvements in the optimal

### Table 3. Comparison of Results for Dedicated vs. Flexible Nurses—Small Instance (5 Patients per Provider) and Large Instance (10 Patients per Provider) for Lognormal Service Times

<table>
<thead>
<tr>
<th></th>
<th>5 patients per provider</th>
<th>10 patients per provider</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dedicated nurses</td>
<td>Flexible nurses</td>
</tr>
<tr>
<td>Mean wait time</td>
<td>11.75</td>
<td>10.35</td>
</tr>
<tr>
<td>Mean idle time</td>
<td>37.12</td>
<td>34.92</td>
</tr>
<tr>
<td>90th percentile of wait time</td>
<td>22.01</td>
<td>23.52</td>
</tr>
<tr>
<td>90th percentile of idle time</td>
<td>60.00</td>
<td>53.55</td>
</tr>
<tr>
<td>Objective function value</td>
<td>82.89</td>
<td>76.57</td>
</tr>
</tbody>
</table>

*Note. Time values expressed in minutes.*

### Table 4. Comparison of Results for Dedicated vs. Flexible Nurses—Large Instance (10 Patients per Provider) and Lognormal Service Times with Regular and Quadrupled Variance

<table>
<thead>
<tr>
<th></th>
<th>Regular service time variance</th>
<th>Quadrupled service time variance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dedicated nurses</td>
<td>Flexible nurses</td>
</tr>
<tr>
<td>Mean wait time</td>
<td>15.57</td>
<td>14.08</td>
</tr>
<tr>
<td>Mean idle time</td>
<td>56.83</td>
<td>54.52</td>
</tr>
<tr>
<td>90th percentile of wait time</td>
<td>34.62</td>
<td>26.96</td>
</tr>
<tr>
<td>90th percentile of idle time</td>
<td>87.00</td>
<td>81.50</td>
</tr>
<tr>
<td>Objective function value</td>
<td>153.23</td>
<td>143.57</td>
</tr>
</tbody>
</table>

*Note. Time values expressed in minutes.*
patient wait and provider idle time. This suggests that it may not be worth the added complexity it entails for patients.

5. Computational Performance

5.1. Effectiveness of Tightened Formulation

For the case of five patients per provider, we evaluate computational performance of the model, with and without tightening constraints. In the model without tightening constraints, we apply the big-M method with a sufficiently large M value of 200.

In evaluating the computational performance of various approaches, we report the optimality gap, which can be defined as the relative gap between the objective of the best integer solution and the objective of the best node remaining generated by CPLEX. Our model is implemented with IBM ILOG Optimization Programming Language using CPLEX 12.6 and run on a Windows 8.1 pro and 64 bit with Intel(R) Core i7-4770 central processing unit @ 3.40 GHz, 3.401 Mhz, and 32 GB random access memory. The solution and its performance (speed and quality) may depend on the particular sample of scenarios selected, which we denote as a replication. This is especially true if the sample is small. For that reason we generate two replications of 1,000 scenarios by randomly sampling from the empirical service time distribution. The model contains 118,002 constraints, 15,000 binary variables, and 10 integer variables. Tables 8 and 9 present the optimality gap for the various models with and without tightening constraints after one-hour and four-hour run times, respectively.

As shown in Tables 8 and 9, the gap significantly decreases when incorporating tight M values and bounds, with the bounds narrowing the optimality gap far more quickly. It is interesting to note that when running the model for four hours, all models produce the same objectives and schedules. However, we cannot confirm the quality of the solution produced by the formulation without any tightening bounds or tight M. Because of the time limit, the search process has not been completed to guarantee optimality; however, the best integer solution has not improved after a certain time. The significant computational effort shows that “one of the incumbents found in the first minutes of the branch and bound process was indeed the best solution that was to be found” (Topaloglu 2006, p. 383). The objectives and schedules obtained by the model satisfy the goal of the study from the practical viewpoint.

Next we investigate the computational performance of the tightened formulation. As shown in Figure 11, at the end of node 0, the gap reaches close to 5.24% in 62 seconds in the first replication and 4.81% in 70 seconds in the second replication. Within 10 minutes, the gap is 1.2% in the first replication and 1.7% in the second replication.

### Table 5. Comparison of Results for Models with vs. without Crossovers—Small Instance (Five Patients per Provider) for Empirical and Lognormal Service Times

<table>
<thead>
<tr>
<th></th>
<th>Empirical distribution</th>
<th>Lognormal distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Crossover</td>
<td>No crossover</td>
</tr>
<tr>
<td>Mean wait time</td>
<td>10.79</td>
<td>11.81</td>
</tr>
<tr>
<td>Mean idle time</td>
<td>35.76</td>
<td>36.98</td>
</tr>
<tr>
<td>90th percentile of wait</td>
<td>25.21</td>
<td>27.40</td>
</tr>
<tr>
<td>90th percentile of idle</td>
<td>56.00</td>
<td>59.50</td>
</tr>
<tr>
<td>Objective function value</td>
<td>78.81</td>
<td>82.79</td>
</tr>
</tbody>
</table>

Note. Time values expressed in minutes.

### Table 6. Comparison of Results for Models with vs. Without Crossovers—Small Instance (5 Patients per Provider) and Large Instance (10 Patients per Provider) for Lognormal Service Times

<table>
<thead>
<tr>
<th></th>
<th>5 patients per provider</th>
<th>10 patients per provider</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Crossover</td>
<td>No crossover</td>
</tr>
<tr>
<td>Mean wait time</td>
<td>10.35</td>
<td>10.56</td>
</tr>
<tr>
<td>Mean idle time</td>
<td>34.92</td>
<td>37.28</td>
</tr>
<tr>
<td>90th percentile of wait</td>
<td>23.52</td>
<td>24.21</td>
</tr>
<tr>
<td>90th percentile of idle</td>
<td>53.55</td>
<td>58.00</td>
</tr>
<tr>
<td>Objective function value</td>
<td>76.57</td>
<td>80.77</td>
</tr>
</tbody>
</table>

Note. Time values expressed in minutes.
The objective after 10 minutes is only 0.03% and 0.2% higher than that after four hours, respectively. Therefore, the tightened formulation leads to a near-optimal solution very quickly.

5.2. Impact of Lower Bound Based on Solving Mutually Exclusive Scenario Subsets

For the medium and large cases (8 and 10 patients per provider, respectively), we apply the lower bounding technique described in Section 3.3. For the medium case we split the 1,000 scenarios of the medium practice into 50 mutually exclusive groups of 20. It takes 3 hours and 45 minutes to solve the 50 groups to create the lower bound, but because that lower bound is very tight, the optimality gap for the original full 1,000-scenario problem decreases down to 8% within 10 minutes, and 2% within 4 hours.

Because of the high computation times, for the large case we use the lower bounding technique with even smaller subsets of scenarios: 100 mutually exclusive subsets with 10 scenarios in each subset. It takes 1 hour and 10 minutes to solve the corresponding 100 subproblems to create a lower bound, but because that lower bound is tight, the optimality gap decreases down to 6.4% within 4 hours (5 hours and 10 minutes total) while solving the global problem.

Thus, we conclude that the lower bounding technique is extremely useful in cutting down the optimality gap for large instances. Without this lower bound, optimality gaps can be as high as 22% for large instances at the end of 4 hours.

When solving instances drawn from various lognormal service time distributions, we find that as the service time variance increases relative to a fixed mean, problem TP gets computationally easier. Again, detailed results can be found in the e-companion to our paper.

6. Extension to Incorporate No-Shows

Until now we have assumed that scheduled patients always show up to their appointments, because the practice that inspired our study has only a 3% patient no-show rate. In this section, we study the performance of our models and suggest scheduling guidelines for various no-show rates. We consider no-show rates ranging from 5% to 30% (Cayirli and Veral 2003), in increments of 5%. The method to model the different no-show rates within our stochastic programming formulation is to randomly place zero-length visit durations with nurse and provider in the data used to generate the scenarios. This approach captures provider idle time exactly but results in an approximation of patient wait times. Although the wait time of all the patients seen by the provider is calculated correctly, the objective function also includes the wait time the patient who did not show up would have experienced. As in previous sections, we optimize appointment times by solving the model with the tightened formulation, but we make sure to allow for double-booking of appointment slots.

Figure 12 displays the schedule under different no-show rates. As expected, the schedule gets packed when no-show rates increase. With a 25% no-show rate, slack is no longer needed in the schedule, even when double-booking the first two patients of one of the providers. The optimal schedule under 30% patient no-shows includes

Table 7. Comparison of Results for Models with vs. Without Crossovers—Large Instance (10 Patients per Provider) for Lognormal Service Times with Regular and Quadrupled Variance

<table>
<thead>
<tr>
<th></th>
<th>Regular service time variance</th>
<th>Quadrupled service time variance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Crossover</td>
<td>No crossover</td>
</tr>
<tr>
<td>Mean wait time</td>
<td>15.38</td>
<td>14.14</td>
</tr>
<tr>
<td>Mean idle time</td>
<td>48.36</td>
<td>56.07</td>
</tr>
<tr>
<td>90th percentile of wait time</td>
<td>30.41</td>
<td>28.77</td>
</tr>
<tr>
<td>90th percentile of idle time</td>
<td>74.55</td>
<td>84.05</td>
</tr>
<tr>
<td>Objective function value</td>
<td>138.93</td>
<td>146.27</td>
</tr>
</tbody>
</table>

Note. Time values expressed in minutes.

The objective after 10 minutes is only 0.03% and 0.2% higher than that after four hours, respectively. Therefore, the tightened formulation leads to a near-optimal solution very quickly.

Table 8. Computational Performance for Models with and Without Tightening Constraints with Allowance of One Hour

<table>
<thead>
<tr>
<th>Gap</th>
<th>Model with large M</th>
<th>Model with tight M</th>
<th>Model with large M and bounds</th>
<th>Model with tight M and bounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>First replication</td>
<td>46.93</td>
<td>14.02</td>
<td>1.46</td>
<td>1.05</td>
</tr>
<tr>
<td>Second replication</td>
<td>59.80</td>
<td>15.59</td>
<td>1.54</td>
<td>1.34</td>
</tr>
</tbody>
</table>

Note. Values are percentages.
one open slot (slack) because both providers are double-booked at the beginning of the session. Double-booking the first two patients of provider 1 is a robust scheduling guideline in the range of 5%–30% no-shows. The double-booking is followed by an open slot for no-show rates in the range of 5%–20%; no slack is necessary under 30% no-shows. It is also interesting to note that the slack position is pushed down, to later in the schedule, as the no-show rates increase. Because of no-shows, wait time and idle time accumulate at a slower pace. In addition, although our formulation allows double-booking any two consecutive patients, the optimal solutions generated only suggest double-booking the very first two appointments.

7. Conclusion

In this paper, we model and solve a challenging outpatient scheduling problem: the team primary care practice. The system can be modeled as a tandem queue: the patients are first seen by a team of nurses in a flexible manner and then seen by their dedicated provider. We restrict our study to the case of two nurses and two providers, which is highly relevant because larger practices often operate in smaller independent teams such as the one studied.

Although an FCFS queueing policy at the second step (provider) is attractive in practice, it results in a significant modeling challenge because patients will crossover and see the provider in a sequence different from that suggested by their appointment times. We develop a stochastic mixed integer program to solve this unique problem and generate insights based on the structure of the optimal schedules.

In particular, we draw the following conclusions:

1. The optimal schedule is staggered, introducing slack for the two providers in different time slots. The later slack is assigned to the provider whose patients are given priority at the nurse step, resulting in a more packed schedule with potentially fewer open appointment slots for that provider.

2. A deterministic model based on average nurse and provider visit times provides a fast, high-quality heuristic for the stochastic team care problem, producing schedules within 2% of optimality for all instances tested under the cost ratio of 4 (provider idle time is weighted 4 times higher than patient wait time) suggested by the practice we collaborated with. When a heavier weight is placed on patient wait time the quality of the deterministic solution deteriorates. For a cost ratio of 1, the optimality gap is approximately 10% for all levels of service time variance tested.

3. As the variance of the service times increases, the benefits of nurse flexibility and accounting for crossovers on wait time and idle time grow. An optimized schedule allowing for nurse flexibility and patient crossovers leads to significantly lower wait time and idle time when the service time variance is high.

4. As the relative value placed on idle versus wait time is decreased, the optimal schedules approach the current practice policy whereby a slack slot is introduced after every patient.

5. The advantages in reduced wait and idle time of increasing the granularity of appointment slots are rather small and do not outweigh the operational disadvantage in difficult-to-remember appointment times for patients.

**Table 9.** Computational Performance for Models with and Without Tightening Constraints with Allowance of Four Hours

<table>
<thead>
<tr>
<th>Gap</th>
<th>Model with large M</th>
<th>Model with tight M</th>
<th>Model with large M and bounds</th>
<th>Model with tight M and bounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>First replication</td>
<td>28.52</td>
<td>10.54</td>
<td>1.27</td>
<td>0.91</td>
</tr>
<tr>
<td>Second replication</td>
<td>32.16</td>
<td>10.74</td>
<td>1.32</td>
<td>1.13</td>
</tr>
</tbody>
</table>

*Note.* Values are percentages.
6. As the no-show rate increases, the optimal schedules get more packed (introduce less slack), the slack slots are scheduled later, toward the end of the time horizon, and double-booking of the first two patients becomes attractive.

Because of the computationally challenging nature of the problem, we developed methods to improve the solution time and optimality gap. Specifically, we generate a strong lower bound by solving the stochastic mixed integer program for exhaustive and mutually exclusive subsets of the full set of scenarios. We also tighten the big-M parameters using the problem structure and generate additional cuts and constraints to close the optimality gap even further. This allowed us to solve realistically sized problems within a reasonable time frame and optimality gap.

Appendix. Proof of Theorems 1 and 2
Proof of Theorem 1. For Constraint (7) to be valid, we must ensure that

$$M_i^s \geq \left( N_{i-1,s}^{\text{max}} - y_{i-1,s}^{\text{finish}} \right)^+ \quad \forall i \in I_s, s \in S$$

where $N_{i-1,s}^{\text{max}}$ is the maximum of the finish times of patients 1 through $i - 2$ with a nurse for that scenario. That is, $M_i^s$ must be an upper bound on the difference in finish times with nurse of the two patients that are seen by a nurse at the time patient $i - 1$ starts service, and it can vary for each patient in the sequence and from scenario to scenario. We consider two cases: In case 1, the finish time of patient $i - 1$ with nurse is greater than or equal to the maximum of the finish times of patient from 1 to $i - 2$ with nurses; and in case 2 it is strictly lower.

Case 1: $N_{i-1,s}^{\text{max}} \leq y_{i-1,s}^{\text{finish}}$
In this case, observe that

a. The appointment time of patient $i - 1$ is at most 30 minutes after that of patient $i - 2$, and thus $X_{i-1} \leq y_{i-2,s}^{\text{start}} + 30$.

b. By definition: $N_{i-1,s}^{\text{max}} = y_{i-2,s}^{\text{start}} + t_i N_{i-2,s}$, and thus $N_{i-1,s}^{\text{max}} - y_{i-2,s}^{\text{start}} \geq y_{i-2,s}^{\text{finish}}$.

c. Combining the two, we get that patient $i - 1$ is available at time:

$$X_{i-1} \leq y_{i-2,s}^{\text{start}} + 30 \leq N_{i-1,s}^{\text{max}} - \tau_i N_{i-2,s}^{\text{max}} + 30.$$

d. A nurse will be available to serve patient $i - 1$ at time $N_{i-1,s}^{\text{max}}$ or earlier.

e. The start time of patient $i - 1$ with the nurse is

$$y_{i-1,s}^{\text{start}} \leq \max\{N_{i-1,s}^{\text{max}} - N_{i-2,s}^{\text{max}} - \tau_i N_{i-2,s}^{\text{max}} + 30\}.$$
Proof of Theorem 2. For the $M^2$ constraints to be valid we must ensure that

$$M^2_{j,s} \geq \left( P_{j,s}^{\text{max}} - y_i + x_{i+2,s} \right)^+$$

where $P_{j,s}^{\text{max}}$ is the maximum of the finish times at the nurse step of patients $1, \ldots, j$ of provider $k$'s panel under scenario $s$. That is, $M^2$ is a bound on the difference of nurse finish times of subsequent patients seen by provider $k$, for $j = 1, 2, \ldots, l_j$ for provider $k$.

Observe that if $y_i = y_i$ then $y_i$ is the $j$th patient in provider $k$'s panel, who is the $i$th patient in the practice.

We again consider two cases: in case 1 the nurse finish time of patient $i + 2$ is greater than or equal to the maximum of the nurse finish times of patients of provider $k$ up to patient $j$; and in case 2 it is lower.

Case 1: $P_{j,s}^{\text{max}} \leq y_i + x_{j+2,s}$

In this case, observe that

a. The appointment time of patient $i + 2$ is at most 30 minutes after that of patient $i$; thus

$$X_{i+2} \leq y_i + x_{i+2,s} + 30$$

b. By definition: $P_{j,s}^{\text{max}} \leq y_i + x_{i+2,s}$ and thus $P_{j,s}^{\text{max}} - y_i + x_{i+2,s} \geq y_i - y_i + x_{i+2,s}$

c. Combining the two, we get that patient $i + 2$ is available at time:

$$X_{i+2} \leq y_i + x_{i+2,s} + 30 \leq P_{j,s}^{\text{max}} - y_i + x_{i+2,s} + 30$$

d. Patient $i + 1$ (from the other provider’s panel) will be seen by a nurse at a time no later than max$(P_{j,s}^{\text{max}}, y_i + x_{i+2,s} + 30)$. This is using that consecutive patients arrive at most 30 minutes apart to the practice.

e. A nurse will be available for patient $i + 2$ at time max$(P_{j,s}^{\text{max}}, y_i + x_{i+2,s} + 30) + t_{i+1 sabe}$ or earlier. This is a bound on the time a nurse will be available if patient $i + 2$ is scheduled to see a nurse right after patient $i + 1$ finishes.

f. The appointment time of patient $i + 2$ is at most 30 minutes after that of patient $i$; thus

$$X_{i+2} \leq y_i + x_{i+2,s} + 30$$

In this case, observe that although patient $i + 2$ has finished with one nurse, say nurse 1 w.l.o.g., the other nurse, nurse 2, is still busy with an earlier patient $r$ from the same provider. The difference between the two can be calculated depending on which patient is still with nurse 2. If patient $r$ is still with nurse 2, it means that patients $r + 1, r + 2, \ldots$, through $i + 1$ were seen by nurse 1, we have that:

a. $y_i^{start} + x_{i+2,s} \geq y_i^{finish} + x_{i+2,s}$

b. $y_i^{finish} + x_{i+2,s} \geq y_i^{finish} + x_{i+2,s}$

c. $y_i^{start} + x_{i+2,s} \geq y_i^{finish}$ because patients are seen in the order of their appointment times, $X_i \leq X_2 \leq \cdots \leq X_{f,j}$

d. Thus, the difference $P_{j,s}^{\text{max}} - y_i + x_{i+2,s}$ is bounded by $\tau_{i+1,s} + t_{i+1,s} + \cdots + t_{i+2,s}$

The maximum in max$_{r=i+1, \ldots, j}$ in provider’s k panel will give us the bound we are looking for in this case.

The overall bound on the difference for both cases then is

$$\max \{\text{Case 1, Case 2} \} = \max \left\{ \nabla_{i+1, \ldots, j \text{ in provider’s k panel}} \{ \nabla_{i+1, \ldots, j \text{ in provider’s k panel}} \} \right\}$$

References


