IISE TRANSACTIONS
https://doi.org/10.1080/24725854.2018.1542544

Extended open shop scheduling with resource constraints: Appointment scheduling for integrated practice units

Pengfei Zhang[a], Jonathan F. Bard[b], Douglas J. Morrice[a], and Karl M. Koenig[c]

[a]McCombs School of Business, The University of Texas, Austin, TX, USA; [b]Cockrell School of Engineering, The University of Texas, Austin, TX, USA; [c]Medical Director of the Integrated Practice Unit for Musculoskeletal Care, Dell Medical School, The University of Texas, Austin, TX, USA

ABSTRACT

An Integrated Practice Unit (IPU) is a new approach to outpatient care in which a co-located multidisciplinary team of clinicians, technicians, and staff provide treatment in a single patient visit. This article presents a new integer programming model for an extended open shop problem with application to clinic appointment scheduling for IPUs. The advantages of the new model are discussed and several valid inequalities are introduced to tighten the linear programming relaxation. The objective of the problem is to minimize a combination of makespan and total job processing time, or in terms of an IPU, to minimize a combination of closing time and total patient waiting time. Feasible solutions are obtained with a two-step heuristic, which also provides a lower bound that is used to judge solution quality. Next, a two-stage stochastic optimization model is presented for a joint pain IPU. The expected value solution is used to generate two different patient arrival templates. Extensive computations are performed to evaluate the solutions obtained with these templates and several others found in the literature. Comparisons with the expected value solution and the wait-and-see solution are also included. For the templates derived from the expected value solution, the results show that the average gap between the feasible solution and lower bound provided by the two-step heuristic is 2% for 14 patients. They also show that either of the two templates derived from the expected value solution is a good candidate for assigning appointment times when either the clinic closing time or the patient waiting time is the more important consideration. Sensitivity analysis confirmed that the optimality gap and clinic statistics are stable for marginal changes in key resources.

ARTICLE HISTORY
Received 3 January 2018
Accepted 23 October 2018

KEYWORDS
Integrated practice units; open shop scheduling; flexible flow shop; stochastic optimization

1. Introduction

The United States spent nearly 18% of its gross domestic product on healthcare in 2015 according to the Centers for Medicare & Medicaid Services (CMS). In 2016, U.S. healthcare spending reached a new peak at $10,348 per person, more than twice the average of other developed countries. Today, it is most common for patients who need multiple consults to travel from one clinic to another to see different providers. Such a provider-centered approach inevitably burdens the patient in the following ways: (i) travel between facilities introduces inconvenience, additional logistics costs and unnecessary administrative costs; (ii) repeated requests for the same information can increase stress and anxiety; (iii) information transfer across clinics often results in inaccurate or incomplete health records downstream; (iv) lack of communication among providers may occasion unnecessary or duplicate tests, and undermine long-term care planning; and (v) the separation of providers reinforces a piecemeal approach that rarely addresses the patient's condition as a whole. To better deliver healthcare services, current healthcare reform is moving towards value-based patient-centered care, seeking better coordination among providers. Many researchers have shown that this approach can improve clinical outcomes while decreasing diagnostic tests and the need for referrals (Stewart *et al.*, 2000; Hanna, 2010).

1.1. Integrated practice units

To put the focus on the needs of the patient, several clinicians and policy analysts have suggested the use of Integrated Practice Units (IPUs) to treat chronic medical conditions such as diabetes, pain, multiple sclerosis, and cardiomyopathy (Porter 2010; Keswani *et al.* 2016). This approach fosters realtime communication among specialists and provides treatment options for the patient across the entire continuum of care for a chronic condition. An added benefit inherent to this model is the continuous learning and improvement of multiple disciplines working together and communicating about each patient. The team learns from every patient, so their expertise improves over time. After a patient enters the IPU and is roomed, the appropriate providers sequentially address the patient's conditions. In some cases, it is appropriate for the patient to see

CONTACT Jonathan Bard jbard@utexas.edu University of Texas at Austin, Operations Research and Industrial Engineering, Austin, TX
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uiie.

different providers in a specific order. For example, in the case of a lower extremity joint pain IPU, the motivating clinic for this article, the patient is first seen by a nurse practitioner who determines whether additional treatment is required. If it is decided that the patient needs to see both a surgeon and a physical therapist, the surgeon comes first. If it is determined that the patient must see a physical therapist and a nutritionist, the order is immaterial.

What is relatively unique about an IPU is that the patient remains in the same room for the duration of his/her visit, and hence is the center of pathways traversed by a variety of providers. This model of care delivery enables the providers to work more closely together in treating their patients, and to focus on using the skills for which they have been trained. The expectation is better outcomes, higher levels of patient satisfaction, and lower patient costs in the long run. What has yet to be determined, though, is whether the efficiency of an IPU will outweigh the higher provider costs that are likely to result from lower provider utilization. To be effective, all providers must be available in the IPU, but not all patients need to see all providers.

While IPUs bring continuity of care and integrated treatment to patients – important factors in patient satisfaction – they also present an operational challenge to the schedulers who must coordinate activities among all providers. In the current system, healthcare delivery is fragmented; patients see their providers at different times and often at different locations. In an IPU, the patients have seamless access to service from different providers in a timely manner; however, this requires better coordination among providers to prevent delays and congestion. Moreover, in the fragmented delivery system, different clinics operate independently and each has its own scheduling system. In an IPU, the schedules of the different providers interact with each other, as the patient needs to see one provider followed by another. As a consequence, clinic scheduling becomes central to the efficiency of the multidisciplinary team, as patient demand and provider capacity have to be strategically matched to ensure timely operations.

One of the most critical issues in managing an IPU is deriving the appointment schedule or template. Ill-conceived templates result in excessive patient waiting time, unacceptably low provider utilization, and costly overtime. The challenge then is to design schedules that jointly balance clinic closing time and total patient waiting time while also taking into account system capacity and system randomness. For a given number of patients, if these two metrics are minimized, then provider and staff idle time should also be minimized. The system randomness derives from two sources. The first is each patient's provider set. These sets are unknown at the time when the appointment is made and are only determined after the patient is seen by the nurse practitioner who conducts an initial examination. The second is the amount of time each patient spends with each provider.

In practice, the coordination of providers in IPUs has many elements of an extended open shop scheduling problem in which each workstation may consist of multiple (identical) machines and some jobs have partially fixed routes. The pure open shop problem has been studied extensively in the combinatorial optimization literature, and is known to be strongly NP-hard (see Pinedo 2016). Its aim is to assign a set of jobs to available machines to minimize one of several objective functions such as makespan, total processing time, or number of late jobs. As the IPU appointment scheduling problem has many similar characteristics as the open shop problem, the models for either are quite similar. In the case of an IPU, the patients can be viewed as jobs and the providers as machines. Clinic performance is measured by patient delay (total processing time) and closing time (makespan). These measures are in conflict so a strategic balance must be struck.

1.2. Research contributions

The purpose of this article is to first present a generic model of the IPU scheduling problem and then to develop a solution methodology for realistic size instances. We focus on two decisions: the number of patients to schedule in each time period and appointment rules. The first decision is intended to fix the appointment template, which specifies how many patients should be scheduled to arrive at the beginning of each time slot. Appointment rules determine which types of patients (new or follow-up) to assign to each time slot. We begin with a deterministic model based on an open shop that takes into account the unique characteristics of an IPU including different types of providers, multiple providers of the same type, fixed and variable patient paths, and patient waiting time limits. An additional consideration is the number of available rooms. Once a patient checks into the clinic, he/she is assigned to a room and remains there until the visit is concluded. This type of resource constraint is not often modeled in open shop scheduling problems where the common restrictions center on labor and machines. It is rare for auxiliary resources, such as rooms, transportation vehicles, and other tooling and equipment to be taken into account. We propose three approaches to modeling such resource constraints. To capture the randomness of provider service times and patient-specific treatments, we show how our generic model can be extended to include these stochastic elements. The approach is demonstrated using data provided by The University of Texas Dell Medical School in Austin.

This research differs from earlier studies in the following ways. To the best of our knowledge, this is the first article to present a generic (stochastic) model for determining appointment templates for a multi-stage, multi-server, resource-constrained clinic where patients remain stationary throughout their visit. Another unique feature of our problem is the order of provider–patient engagement. Existing studies usually assume that the patient sees the providers in a fixed sequence if there is more than one, whereas in our case, the order is only fixed for some providers while remaining flexible for the others. Thus, the IPU scheduling problem is really a combination of a flexible flow shop and an open shop with auxiliary resource constraints (see Pinedo

2016). The two-step method proposed to find solutions is sufficiently general to be used to help solve similar coordinated appointment scheduling problems arising from other applications.

The remainder of this article is structured as follows. In Section 2, we provide a literature review of the most relevant work on open shop scheduling and healthcare appointment scheduling. In Section 3, we present our new model for the extended open shop scheduling problem, analyze its features, and introduce several valid inequalities that were seen to speed convergence. We also describe our two-step solution method. The random components of the problem are introduced in Section 4, where we present a two-stage stochastic optimization model and define what we mean by the expected value solution and the wait-and-see solution. In Section 5, we examine the relative performance of two templates derived from the expected value solution and two found in the literature. Extensive testing is done to compare IPU metrics across all templates and to evaluate the quality of the lower bound obtained from the two-step method. The results indicate that the average gap for the two-step method is always less than 5% for the wait-and-see problem, and less than 2% for the four appointment templates that we investigated.

We also observed that the two templates derived from the expected value solution are good candidates for setting appointments. One template emphasizes the clinic closing time objective by scheduling patients to arrive relatively earlier in the day. The second template emphasizes the patient waiting time objective by scheduling patients to arrive later in the day. Lastly, the results show that our appointment rules are helpful when scheduling the different types of patients. For example, we found that it is best to schedule follow-up patients, who generally have shorter service times, to arrive when there is high patient flow. This helps to relieve or avoid congestion when the number of patients is fixed over the day. We conclude with some managerial insights and some suggestions for future research in Section 6.

2. Literature review

Variations of job shop problems have been studied extensively and have a wide range of applications. One common example is the open shop problem in which a set of jobs is to be processed through multiple stations in an arbitrary order, as is partially the case in an IPU. Bhat *et al.* (2000) modeled the communication scheduling problem as an open job shop whereas Liaw (2000) proposed a hybrid genetic algorithm that incorporated tabu search as part of the solution methodology. Noori-Darvish *et al.* (2012) developed a bi-objective mixed-integer linear programming (MILP) model for an open shop scheduling problem with sequence-dependent setup times, and applied an interactive fuzzy programming approach to find solutions. Our clinic scheduling problem can be modeled as an extended open shop, where "extended" means multiple, parallel machines, fixed and arbitrary job processing paths, and auxiliary resource constraints.

Scheduling problems in healthcare often have special features that distinguish them from problems arising in other industries. Their unique nature brings additional challenges. For example, Gupta and Denton (2008) noted that in healthcare applications there exists less flexibility because patients may have a preference for a specific provider or appointment time. Moreover, urgent patient needs must be accommodated immediately, and in some cases, price cannot be used to modulate patient demand. With respect to outpatient scheduling, a wide variety of approaches have been investigated, but few have been implemented in practice. In the remainder of this section, we provide a literature review of healthcare scheduling problems with different system structures. We also provide a review of the different solution methods with an emphasis on stochastic programming approaches.

2.1. Healthcare systems with different pathway structures

There have been numerous studies on scheduling in healthcare over last several decades, as highlighted by Cayirli and Veral (2003) and Gupta and Denton (2008). Most of the early work focused on single-station appointment scheduling. More recently, the scope has expanded to include multi-stage, multi-server applications as discussed by Ahmadi-Javid *et al.* (2017) and Leeftink *et al.* (2018). Based on the features of our problem, the most relevant studies can be grouped into two categories: multi-stage models and multi-server models. Each is reviewed below.

In multi-stage clinic scheduling, different provider types are involved. This makes the problem complicated because a patient can be referred from one provider to another for different treatment, which leads to uncertainty in the patient flow. Azadeh *et al.* (2015) formulated a semi-online patient scheduling problem as a MILP, and developed a genetic algorithm to find solutions. In their problem, the patients require different types of tests and the use of a variety of laboratory equipment. Castro and Petrovic (2012) studied a scheduling problem in which patients need to go through an ordered sequence of examinations. They formulated the problem as a three-objective mathematical program, and solved it with a dispatching rule. Pérez *et al.* (2013) investigated a stochastic online scheduling problem for nuclear medicine clinics where the patients need to go through multiple steps. In the study, the sequence of the steps is fixed, and multiple resources are required at each step. Kazemian *et al.* (2017) developed a simulation model to coordinate clinic and surgery appointments with the objective of reducing the indirect waiting time of patients and limiting operating room overtime. Their strategy was to choose appointment days for patients rather than setting daily arrival times. Different from our work, these studies are either limited to a single server at each stage or they do not include room constraints.

Problems get more challenging when there is more than one provider of each type, giving rise to the multi-server clinic scheduling problem. Gupta and Wang (2008) modeled

an appointment booking problem as a Markov decision process and proposed heuristics to find solutions. They also developed lower and upper bounds on the optimal solution, which were shown to speed convergence. Both single- and multiple-physician clinics were analyzed, but in either case, only single-stage scheduling was applicable. Parizi and Ghate (2016) went a step further and purposed a Markov decision process for a multi-class, multi-resource clinic scheduling problem, whereas Qu *et al.* (2013) developed a weekly scheduling template for a multiple-provider outpatient clinic. In their problem, providers in separate sessions have separate appointment schedules, whereas in our study, all providers are in the same clinic working with a single appointment schedule.

2.2. Solution methods for healthcare scheduling problems

Dynamic programming has been a popular tool for modeling the clinic scheduling problem. For example, Truong (2015) considered the problem in which two types of patients are adaptively given appointments over several days. Chakraborty *et al.* (2010) used a dynamic programming tree to investigate clinic scheduling with general service time distributions, where the patients sequentially request appointments. Simulation is perhaps the most versatile tool, since it is able to handle most complexities surrounding patient flow and uncertainty. Wang *et al.* (2018) solved a two-server scheduling problem using simulation-based optimization. Cayirli *et al.* (2006) developed a simulation model to analyze appointment scheduling for ambulatory care and investigated patient sequence rules based on patient class. Similarly, Bard *et al.* (2016) used discrete event simulation to investigate the performance of the family health center associated with the University of Texas Medical School in San Antonio. Their objective was to obtain a better understanding of patient flow and to evaluate changes to current scheduling rules and operating procedures. As part of the study, they examined a variety of scenarios related to appointment scheduling and managing early and late arrivals.

Robust optimization is a relatively new approach to scheduling patients and resources in healthcare facilities. Denton *et al.* (2010) built a robust optimization model to study the allocation of operating rooms to surgical specialties in the face of insufficient data. Rachuba and Werners (2014) applied the robust approach to a hospital surgery scheduling problem in an effort to avoid frequent rescheduling due to random requests and cancellations. Similarly, Mannino *et al.* (2012) presented a light robustness procedure to handle random fluctuations in demand when constructing cyclic master surgery schedules. In their procedure, parameter values lie in an uncertainty set, but solutions are not required to satisfy all possible realizations. Instead, soft constraints are introduced for each parameter and violations are penalized in the model's objective function.

Another common approach to modeling uncertainty is stochastic programming. Mancilla and Storer (2012) considered a stochastic appointment scheduling problem and proposed a new sequencing algorithm based on Benders decomposition to find solutions. Oh *et al.* (2013) used a stochastic integer programming model to schedule patient appointments in primary care facilities and developed scheduling guidelines. Integral to their work is (i) an empirically based classification scheme to distinguish chronic and acute conditions, (ii) the ability to coordinate patient and provider interactions, and (iii) the introduction of slack in the schedule to accommodate the effects of service time variability. Kong *et al.* (2013) investigated an outpatient clinic appointment scheduling problem with a single physician and proposed a convex conic programming approach to find solutions. Berg *et al.* (2014) considered a profit-maximization scheduling problem in the presence of patient no-shows and random procedure times. They modeled the problem as a two-stage stochastic mixed-integer program and proposed several methods to find solutions including two decomposition approaches and a heuristic.

Chen and Robinson (2014) formulated a clinic scheduling problem with both routine patients and last-minute patients as a stochastic linear program. They derived optimal sequencing rules while accounting for random no-shows and call-ins. Erdogan and Denton (2013) proposed a multi-stage stochastic linear program in which each stage is defined to coincide with the time a patient calls to request an appointment. Different from the formulations in these studies, our two-stage optimization model accounts for resources shared among patients and co-located providers who see patients in a partially fixed and partially random order.

3. Deterministic model

As noted in Section 1.1, it is critical to consider uncertainty when designing appointment templates for IPUs. The foundation of our approach is a stochastic optimization model whose solution relies heavily on efficiently solving a deterministic version of an Extended Open Shop Scheduling (EOSS) problem. In Section 3.1, we present our EOSS model that includes parallel machines at each station. After describing the formulation, we highlight its unique features in Section 3.2 and offer some tightening constraints designed to reduce the computational burden. To make the discussion concrete, the focus is on clinic scheduling, but with the understanding that the model is generally applicable to most open shop problems. Next, in Section 3.3 the formulation for the room constraints is presented. These constraints can readily handle similar resources such as vehicles, jigs, tooling, and auxiliary personnel. In Section 3.4 we specialize the open shop model to an IPU and impose additional restrictions that better reflect operational considerations. Finally, in Section 3.5 we propose a two-step heuristic to obtain upper and lower bounds on the optimal schedule.

3.1. Extended model for open shop scheduling

We first study the general minimum makespan extended open shop scheduling problem with a secondary objective of

$$\underbrace{\text{A, B, C}}_{\text{first set}}, \underbrace{\text{A, B, C}}_{\text{second set}}, \text{A, } \ldots \text{ , } \underbrace{\text{A, B, C}}_{m^k th \text{ set}}$$

Figure 1. Patient positions for provider type k with three providers.

minimizing the total time that jobs spend in the system. The presentation reflects clinic appointment scheduling rather than job shop scheduling. In the developments, we make use of the following notation.

Indices and sets

- i, j index for patients
- k, l index for providers or provider types
- m index for position in the sequence of patients who see a particular provide type
- o origin (and destination) index for all patients and all providers
- J set of patients
- K set of provider types
- $J(k)$ set of patients who see type k provider
- $K(j)$ set of provider types that patient j needs to see

Data and parameters

- adm^k time (hours) required for a type k provider to perform administrative functions such as entering data into the electronic medical records system after seeing each patient
- LT_m^k lower bound on $(m+1)$st patient's starting time with a type k provider. (When there is only one provider of type k, LT_m^k equals the sum of the m smallest service times of provider k's patients. It is also the lower bound on the time interval between any two patients who are separated by $m-1$ other patients.)
- m^k total number of patients that type k providers are to see
- n^k number of type k providers
- s_j^k service time required for a type k provider to treat patient j (hours)
- ϵ ratio of the predetermined waiting time that a patient can spend in clinic to the patient's total service time
- $S_j(\epsilon)$ upper limit on the amount of time that patient j is allowed to spend in the clinic, or equivalently, the total service time plus upper limit on waiting time of patient j; that is $(1+\epsilon)\cdot\sum_{k\in K(j)} s_j^k$
- T_{max} upper bound on clinic closing time

Decision variables

- t_m^k start time of the patient in the mth position in the schedule of type k providers
- x_{jm}^k 1 if patient j is in the mth position in the sequence of patients who see a type k provider, 0 otherwise
- ST_j^k time when a type k provider starts seeing patient j
- y_j^{kl} 1 if $ST_j^k + s_j^k \le ST_j^l$, which means that a type k provider must finish his/her visit with patient j before a type l provider can start seeing patient j;
  0 if $ST_j^k \ge ST_j^l + s_j^l$, which means that a type k provider can start seeing patient j no earlier than a type l provider finishes his/her visit with patient j

Accounting variables

- T clinic closing time
- T_j^1 time when patient j is seen by his/her first provider
- T_j^2 time when patient j finishes being seen by his/her last provider

For the clinic scheduling problem, we are given a set J of $|J|$ patients and a set K of $|K|$ provider types. For each $k \in K$ there are n^k providers. Different providers of the same type can perform the same tasks. Each patient $j \in J$ needs to be seen by a subset of providers, denoted by $K(j)$. The service time for patient j when treated by a type k provider is s_j^k. As in the general open shop scheduling model, there is no restriction on the order in which providers can see patients.

Each patient is visited by one provider at a time and cannot be preempted once service begins. When the provider finishes treating a patient, she/he documents the episode. This requires a moderate amount of administrative time, but does not affect the patient who can be seen immediately by another provider. The objective is to minimize a weighted combination of the makespan (clinic closing time) and the patients' total time in clinic. The makespan is our primary concern, and in the implementation, is assigned a much larger weight than the total time patients spend in the facility.

To simplify the presentation, first consider the case where $n^k = 1$ for all $k \in K$, where the m^k patients to be seen by the type k provider are indexed by m (i.e., $m = 1, 2, \ldots, m^k$). The decision variable x_{jm}^k is associated with patient $j \in J(k)$ and takes the value of 1 if patient j is in position m in provider k's schedule, and 0 otherwise. The benefit of this indexing scheme is that if a position has a lower/higher index, then the starting time associated with this position should also be lower/higher. Accordingly, the position index can be used to calculate lower and upper bounds on the starting time of the corresponding patient.

Now consider the case where $n^k > 1$. For any k, a corresponding provider can see at most m^k patients. Therefore, we need at most $n^k \cdot m^k$ binary x-variables for each $j \in J(k)$ to determine which of the m^k providers treats patient j, as well as the order in which patients are seen. To help formulate the constraints, we put each provider's patient positions into different sets. Figure 1 depicts an example with three providers A, B, and C of the same type. In the model, there are $3 \cdot m^k$ positions indexed as $1, 2, \ldots, 3 \cdot m^k$, where each position is marked as A, B, or C. The patients who are assigned the positions marked with an A (B or C), will be seen by provider A (B or C, respectively). In the example, provider A's patients will be in positions 1, 4, 7, … Since we have m^k patients and $3 \cdot m^k$ positions, only m^k positions will be filled by the patients in a solution; the remaining $2 \cdot m^k$ positions will be empty.

For the general case with n^k type k providers and m^k patients, we have m^k sets, with each set containing n^k positions. The first patient in each set is seen by the first type k provider, the second patient is seen by the second type k provider, and so on. The nth patient in the mth set is the mth patient seen by the nth type k provider. In a solution, only m^k out of the $n^k \cdot m^k$ positions will be occupied. For

provider type k, the binary variable x^k_{jm} specifies which position patient j takes, and according to the indexing scheme, the value of m determines which provider the patient sees. In a preprocessing step it is possible to eliminate a large number of the m^k variables associated with type k providers when $n^k>1$. This is a direct consequence of the following assumption concerning provider-patient assignments.

In the model, we assume without loss of generality that the number of patients assigned to providers of the same type is non-increasing. If there are three type k providers, for example, and 21 ($= n^k$) patients, then the first provider can see up to 21 patients, the second provider can see a maximum of 10 patients, and the third provider can see a maximum of seven patients. A second benefit of the position indexing scheme is that it allows for the implementation of this ordering rule in a straightforward manner.

The model for the EOSS problem is as follows:

$$\min\ \alpha_1 \cdot T + \alpha_2 \cdot \sum_{j\in J}\left(T^2_j - T^1_j\right) \tag{1a}$$

$$s.t.\ \sum_{1\le m\le n^k\cdot m^k} x^k_{jm} = 1, \forall\ j\in J,\ \ k\in K(j) \tag{1b}$$

$$\sum_{j\in J(k)} x^k_{jm} \le 1, m = 1,\ldots,n^k\cdot m^k, \forall\ k\in K \tag{1c}$$

$$\sum_{j\in J(k)} x^k_{jm} \le \sum_{j\in J(k)} x^k_{j,m-n^k},\ m = n^k+1,\ldots,n^k\cdot m^k, \forall\ k\in K \tag{1d}$$

$$\sum_{j\in J(k)} x^k_{jm} \ge \sum_{j\in J(k)} x^k_{j,m+1},$$

$$m\in\left\{1,2,\ldots,m^k\cdot n^k\right\}\setminus\left\{n^k,2\cdot n^k,\ldots,m^k\cdot n^k\right\}, \forall\ k\in K \tag{1e}$$

$$t^k_m - t^k_{m-n^k} \ge \sum_{j\in J(k)} x^k_{j,m-n^k}\cdot\left(s^k_j + adm^k\right),$$
$$m = n^k+1,\ldots,n^k\cdot m^k, \forall\ k\in K \tag{1f}$$

$$y^{kl}_j + y^{lk}_j = 1, \forall\ k\ne l,k,l\in K(j) \tag{1g}$$

$$ST^l_j \ge ST^k_j + s^k_j - \left(1-y^{kl}_j\right)\cdot S_j(\epsilon),$$
$$\forall\ j\in J, \forall\ k\ne l,k,l\in K(j) \tag{1h}$$

$$ST^k_j \le t^k_{m\cdot n^k+n} + \left(1-x^k_{j,m\cdot n^k+n}\right)\cdot T_{max},$$

$$m = 0,\ldots,m^k-1, n = 1,\ldots,n^k, \forall j\in J,k\in K(j) \tag{1i}$$

$$ST^k_j \ge t^k_{m\cdot n^k+n} - \left(1-x^k_{j,m\cdot n^k+n}\right)\cdot T_{max},$$

$$m = 0,\ldots,m^k-1, n = 1,\ldots,n^k, \forall j\in J,k\in K(j) \tag{1j}$$

$$T^1_j \le ST^k_j, \forall\ j\in J,k\in K(j) \tag{1k}$$

$$T^2_j \ge ST^k_j + s^k_j, \forall\ j\in J,k\in K(j) \tag{1l}$$

$$T^2_j - T^1_j \le S_j(\epsilon), \forall\ j\in J \tag{1m}$$

$$t^k_{n^k\cdot m^k-n} + \sum_{j\in J(k)} x^k_{j,n^k\cdot m^k-n}\cdot\left(s^k_j + adm^k\right) \le T,$$
$$n = 0,1,\ldots,n^k-1, \forall\ k\in K \tag{1n}$$

$$x^k_{jm}, y^{kl}_j \in \{0,1\}, T, t^k_m, ST^k_j, T^1_j, T^2_j \ge 0, \forall\ i,j\in J,$$
$$m = 1,\ldots,n^k\cdot m^k, k\ne l,k,l\in K \tag{1o}$$

The objective function (1a) minimizes the weighted sum of the clinic closing time and the total time patients spend in treatment. The weights α_1 and α_2 should be chosen to reflect the relative importance of each term. In the application, the first term dominates the second, which means that the closing time should be made as small before minimizing the total time in the system. To meet this objective, we set $\alpha_1 \gg \alpha_2$.

Constraints (1b) ensure that every patient j will be seen by exactly one provider of each type in his/her provider set $K(j)$. Note that constraint (1b) is a collection of mutually disjoint special ordered set constraints. In each constraint associated with the (j, k) pair, only one x variable will be one and all others zero. Exploiting this structure in the implementation greatly reduced the computational effort.

Constraints (1c) guarantee that every position in provider type k's schedule is assigned to at most one patient. Constraints (1d) ensure that for each type k provider, positions are assigned in increasing order, starting with 1 and going up to $n^k\cdot m^k$. When $n^k = 1$, all m^k positions will be filled. When $m^k>1$, each provider has m^k available positions, but not all of them will be assigned. Although it seems that this could result in multiple optimal solutions, because the positions are assigned in numerical order this will never be the case. Constraints (1e) specify that if there is more than one provider of type k, then the first provider is always assigned at least as many patients as the second, the second at least as many as the third, and so on. This rule also prevents multiple optimal solutions and has the added benefit of removing symmetry among providers of the same type.

Constraints (1f) specify that for a provider of type k, every patient assigned to his/her needs to be separated in time by at least the service time of the patient in the prior position plus the administrative time (there are no constraints for the first n^k positions because they are occupied by the first patient of the n^k providers). This ensures that providers have enough time between two successive patients. Constraints (1g) are written only for those patients who are to be seen by providers k and l, and enforce the condition that the visits take place in sequence. Constraints (1h) ensure that a provider can only start a visit with a patient after the prior provider finishes with the patient.

Constraints (1i) and (1j) define patient j's starting time with each provider type while constraints (1k) ensure that the clinic visit for patient j begins no later than the time when he sees any of his providers. Constraints (1l) guarantee that the ending time of patient j's visit is no earlier than the time when he sees any of his/her providers plus the corresponding service time. Constraints (1m) limit the total time patient j spends in the clinic (total service time plus total waiting time) to be no greater than a threshold $S_j(\epsilon)$ proportional to his/her total service time. Although the second term in the objective function is aimed at minimizing total clinic time, constraints (1m) are not redundant. Without these constraints, some patients may spend an excessive amount of time in the clinic – a result that we wish to avoid.

Constraints (1n) indirectly define the clinic closing time by restricting it to be no earlier than the ending times of all providers. Alternatively, we could have defined the closing time as the time when the last patient leaves, but in the Linear Programming (LP) relaxation, this value is much smaller than the providers' ending times due to the weakness of constraints (1i) and (1j). Using the proposed definition led to tighter LP relaxations and shorter runtimes. Finally, all variables are defined in constraints (1o).

3.2. Model analysis and improvement

In this section, we investigate some of the characteristics of model (1a) - (1o). First, we show how to use the index information associated with each position to improve the formulation. Next, we show how the LP relaxation can be tightened.

3.2.1. Index information and valid inequalities

The index information for two patients seen by the same provider indicates their relative order. Consider provider type k with $n^k = 1$ and m^k patients. The index of the first patient position is 1 and all other positions for that provider have a later starting time. Given that the positions are ordered, and any two successive positions are separated by the first patient's service time plus the provider's administrative time, we can derive lower and upper bounds on the starting time of each position using its index. For example, the lower bound on the starting time of the second patient is the smallest service time of all patients that are seen by provider k plus his/her administrative time, which is denoted by LT_1^k. The upper bound on the starting time of the last patient position is T_{max} minus the smallest service time of all patients seen by provider k plus his/her administrative time, denoted by $T_{max}-LT_1^k$. Generally, for provider k with $n^k = 1$, the lower bound on the starting time of provider k's patient in position m is LT_{m-1}^k and the upper bound is $T_{max}-LT_{m^k-m+1}^k$.

These bounds allow us to strengthen constraints (1i) and (1j). In the LP relaxation of model (1), (1i) and (1j) are weak constraints, due to the need to make T_{max} sufficiently large to avoid cutting off any feasible solutions. As a consequence, the relaxed feasible region is too large for the branch-and-bound approach to be effective for instances of realistic size. It will be seen, however, that replacing constraints (1i) with constraints (2a) and (2b), and constraints (1j) with constraints (2c) and (2d) provides a tighter LP relaxation. Note that constraints (2a) and (2c) are for $n^k = 1$, and constraints (2b) and (2d) are for $n^k>1$:

$$ST_j^k \le t_{m+n}^k - \sum_{m'\le m-1} LT_{m-m'}^k \cdot x_{j,m'+n}^k + \sum_{m'\ge m+1} \left(T_{max} - LT_{m^k-m'+m}^k\right) \cdot x_{j,m'+n}^k,$$
$$m=0,\ldots,m^k-1, n=n^k=1, \forall\ j\in J, \forall\ k\in K(j) \quad (2a)$$

$$ST_j^k \le t_{m\cdot n^k+n}^k - \sum_{m'\le m-1} LT_{m-m'}^k \cdot x_{j,m'\cdot n^k+n}^k + \left(1-\sum_{m'\le m} x_{j,m'\cdot n^k+n}^k\right)\cdot\left(T_{max}-LT_{m+1}^k\right),$$
$$m=0,\ldots,m^k-1, n=1,\ldots,n^k, n^k>1, \forall\ j\in J, \forall\ k\in K(j) \quad (2b)$$

$$ST_j^k \ge t_{m+n}^k - \sum_{m'\le m-1} \left(T_{max} - LT_{m^k-m+m'}^k\right) \cdot x_{j,m'+n}^k + \sum_{m'\ge m+1} LT_{m'-m}^k \cdot x_{j,m'+n}^k,$$
$$m=0,\ldots,m^k-1, n=n^k=1, \forall\ j\in J, \forall\ k\in K(j) \quad (2c)$$

$$ST_j^k \ge t_{m\cdot n^k+n}^k - \left(1-\sum_{m'\ge m} x_{j,m'\cdot n^k+n}^k\right)\cdot\left(T_{max}-LT_{m^k-m-1}^k\right) + \sum_{m'\ge m+1} LT_{m'-m}^k \cdot x_{j,m'\cdot n^k+n}^k,$$
$$m=0,\ldots,m^k-1, n=1,\ldots,n^k, n^k>1, \forall\ j\in J, \forall\ k\in K(j) \quad (2d)$$

Proposition 1. *Collectively, constraints (2a) and (2b) [constraints (2c) and (2d)] are stronger than their counterparts constraints (1i) [constraints (1j)].*

Proof. See Appendix A.1.

The inequalities in the proof show the tightness of the improved constraints (2) given their equivalence to the original two constraints (1i) and (1j). As noted, the index formulation is a unique feature of our model and is useful in tightening constraints and breaking symmetry. These advantages are not available with the more traditional routing formulation, in which the subscripts on the x variables represent the immediate sequence of two entities, such as vehicles, jobs or patients. In our computational testing, we found that the tightened constraints greatly reduced runtimes.

3.2.2. LP relaxation

Tight LP relaxations of MILPs are essential for computational efficiency. In model (1), this is partially achieved with constraints (1f), which enforce a minimum separation time between patients who are on the schedule of the same provider. To see this, we sum constraints (1f) for a single type k provider. Assume that $n^k = 1$ and denote provider k's ending time by $t_{m^k+1}^k$. This leads to

$$\begin{aligned} t_{m^k+1}^k - t_1^k = \sum_{m=2}^{m^k+1} t_m^k - t_{m-1}^k &\ge \sum_{m=2}^{m^k+1} \sum_{j\in J(k)} x_{j,m-1}^k \cdot \left(s_j^k + adm^k\right) \\ &\ge \sum_{j\in J(k)} \left(\sum_{m=2}^{m^k+1} x_{j,m-1}^k\right) \cdot \left(s_j^k + adm^k\right) \\ &= \sum_{j\in J(k)} \left(s_j^k + adm^k\right) \end{aligned}$$

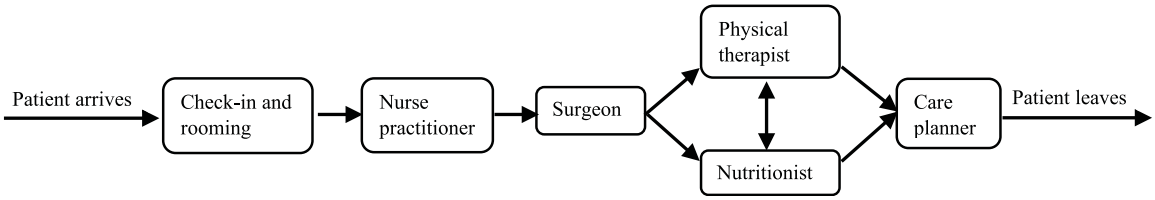

Figure 2. Patient paths in joint pain IPU.

which shows that a provider's ending time and starting time are separated by at least his/her patient's total service time and administrative time. Considering that our primary objective is to minimize the clinic's closing time, which is closely related to providers' ending times, we found empirically that constraint (1f) works in conjunction with constraints (1a) to reduce the computational effort during the branch-and-bound procedure. Network and routing models typically use the Miller-Tucker-Zemlin constraints for the same purpose as constraints (1f), but those constraints include a term equivalent to T_{max} to ensure redundancy when necessary (see Miller *et al.* (1960)). Such formulations are known to provide weak LP relaxations, and proved to be ineffective when trying to solve the stochastic version of the IPU scheduling problem.

3.3. Room constraints

In this section, we present our model for the room constraints. Recall that before a patient can be seen by a provider, he/she is assigned to one of R rooms and remains there until all provider visits are completed. At that point, the room is released and available for the next patient to occupy. When all rooms are in use, arriving patients must wait.

We proposed and tested three methods that equivalently limited the use of rooms to the number available without allowing patients to overlap in the same room. One method may be better than the others, depending on the specific problem. For example, when the number of providers is increased or decreased, the relative performance of the three methods also changes. The most efficient method for our IPU scheduling problem is based on network flow and is presented below. The other two methods are outlined in Appendix B.

Network method. The key variables in this approach are T_j^1 and T_j^2, for all $j \in J$, which appear in constraints (1k) - (1m). Now define a new variable z_{ij} to be 1 if patients i and j use the same room in immediate succession, and 0 otherwise. Let $N = J \cup \{o\}$ be a set of nodes in a network that models patient flow through the clinic, where o is a dummy source/sink node. Between every two nodes in N, we introduce an undirected edge with lower bound 0 and upper bound 1. At the source node, we set the outflow and inflow to be R, and at the patient nodes we set the outflow and inflow to be 1. The patient nodes that receive inflow from the source node correspond to the patients who are the first to use a room. The other flows correspond to the order in which the patients are assigned to rooms.

Let z_{ij} be the flow from node i to node j, for all $i \neq j \in N$. The constraints for room requirement are as follows:

$$\sum_{j \neq i, j \in J \cup \{o\}} z_{ij} = \begin{cases} R, & i = o \\ 1, & i \in J \end{cases} \tag{3a}$$

$$\sum_{i \neq j, i \in J \cup \{o\}} z_{ij} = \begin{cases} R, & j = o \\ 1, & j \in J \end{cases} \tag{3b}$$

$$T_j^1 \geq T_i^2 - (1 - z_{ij}) \cdot T_{max}, \forall\ i \neq j \in J \tag{3c}$$

$$\sum_{m=1}^{m^k} m \cdot x_{jm}^k \geq \sum_{m=1}^{m^k} m \cdot x_{im}^k + 1 - (1 - z_{ij}) \cdot m^k,$$

$$\forall\ i \neq j \in J(k), \forall\ k \in \{k : n^k = 1\} \tag{3d}$$

$$z_{ij} \in \{0, 1\}, \forall\ i, j \in J \cup \{o\} \tag{3e}$$

Constraints (3a) and (3b) specify the outflow and inflow at the nodes, respectively, and together preserve flow balance. Constraints (3c) guarantee that a patient's starting time is no earlier than his/her immediate predecessor's ending time. Constraints (3d) are useful cuts, which state that if patient i leaves his/her room earlier than patient j enters the room, then patient i's position index should be smaller than patient j's position index for any provider who is the only provider of his/her type. The difference must be at least one. Constraints (3e) define the variables.

3.4. Application to joint pain IPU

In this section, we adapt the EOSS model (1) to the joint pain IPU at the Dell Medical School. Provider types include nurse practitioners, surgeons, physical therapists, nutritionists and care planners. The clinic currently operates with two nurse practitioners and one each of the other four provider types. As shown in Figure 2, after self check-in and rooming, every patient is first seen by a nurse practitioner. Depending on the chief complaint, the patient may be seen by one or more of the next three providers. If the patient requires a consult with the surgeon, this takes place immediately after the nurse practitioner. The physical therapist and nutritionist can be seen in any order. Finally, every patient must meet with the care planner at the end of the visit. After a provider finishes with a patient, the next provider can enter the room immediately but the former provider must complete a small number of administrative tasks (e.g.,

writing prescriptions) before moving on to his/her next patient.

In the joint pain IPU, seven exam rooms are available for treatment and consultation. Once assigned to a room, the patient remains there until his/her visit with the care planner ends and he/she departs.

3.4.1. Discrete-time arrival

If there are no other constraints on arrival times, then patient j's appointment time will be T_j^1 minus the time for check-in and rooming. In practice, however, clinic appointment times are assigned at fixed intervals rather than continuously throughout the day as the solution to model (1) would indicate since t_m^k is a continuous variable. For example, if the clinic opens at 8:00 am and we use a 15-minute interval, then patients can be scheduled at 8:00, 8:15, 8:30, ... Assume that each patient spends s_0 minutes on check-in and rooming. Let τ be the minimum time between scheduled appointments and let q be the index for arrival time points. Also, let $x_{jq}^{arr1} = 1$ if patient j arrives at the qth time point (multiple of τ) and sees the first nurse practitioner, and 0 otherwise. Let $x_{jq}^{arr2} = 1$ if patient j arrives at the qth time point and sees the second nurse practitioner, and 0 otherwise. We define a new variable n_q^{arr} to represent the schedule template such that $n_q^{arr} = \sum_j (x_{jq}^{arr1} + x_{jq}^{arr2})$ indicates the total number of patients who arrive at time point q. The following constraints are needed for the discrete-time arrival requirement (for convenience, it is assumed that T_{max} is an integral multiple of τ):

$$\sum_{q=0}^{T_{max}/\tau-1} \left(x_{jq}^{arr1} + x_{jq}^{arr2}\right) = 1, \forall\ j \in J \tag{4a}$$

$$T_j^1 \geq s_0 + \sum_{q=0}^{T_{max}/\tau-1} \left(x_{jq}^{arr1} + x_{jq}^{arr2}\right) \cdot q \cdot \tau, \forall\ j \in J \tag{4b}$$

$$n_q^{arr} = \sum_{j \in J} \left(x_{jq}^{arr1} + x_{jq}^{arr2}\right), q = 0, 1, ..., T_{max}/\tau - 1 \tag{4c}$$

$$x_{jq}^{arr1}, x_{jq}^{arr2} \in \{0, 1\}, \quad n_q^{arr} \in \{0, 1, 2\}, \forall\ j \in J,$$

$$q = 0, 1, ..., T_{max}/\tau - 1 \tag{4d}$$

Constraints (4a) ensure that each patient arrives at the clinic at one of the T_{max}/τ time points. Constraints (4b) guarantee that each patient j is checked in and roomed before being seen by his/her first provider. Constraints (4c) determine the number of patients who arrive at each time point. Constraints (4d) define the variables, where for practical purposes the maximum number of patients who are permitted to arrive at any time point is limited to two. When this bound is relaxed, we found it rare that more than two patients are assigned the same appointment time. As our ultimate goal is to derive appointment templates that are near-optimal for a large number of scenarios with both stochastic service times and patient pathways, a handful of violations will have a negligible effect on the results.

3.4.2. Valid inequalities – lower bounds

The joint pain IPU treats two types or groups of general patients: new and follow-up. New patients usually require longer service times with providers than follow-ups. It is assumed that the ratio of the two patient types is an input parameter. One decision that the model makes is the ordering of the patient types. When a patient calls to schedule a visit, it is known whether he/she is a new or follow-up patient. Therefore, the arrival time can be set based on one of several rules, such as "all follow-ups at the end of the session." Other information about the patient, such as which providers he/she will see and their service times, is not known when the appointment is made. That is, the patient routing is determined after the nurse practitioner encounter during which a diagnosis is made.

Since every patient is assumed to spend the same amount of time for check-in and rooming, they see the nurse practitioner in a first-come, first-served order. This allows us to calculate a lower bound on each patient's starting time with the nurse practitioner. Using similar reasoning, if patient i starts no later than patient j, and there are other patients who start no later than patient j but no earlier than patient i, we can also find a lower bound on the time interval between patient i and patient j's starting time with the nurse practitioner.

Specifically, let A_j be the set of patients of the same type as patient j whose visit with the nurse practitioners starts no later than patient j's, excluding j. Let $M(A_j, n)$ be the sum of the n largest service times with the nurse practitioner of the patients who belong to set A_j. This leads to the following proposition which provides a lower bound on the patients' starting times with the nurse practitioner.

Proposition 2. (Separation Proposition). *If patients j_1 and j_2 are of the same type, and patient j_1's visit with a nurse practitioner begins no later than patient j_2's, then*

$$ST_{j_2}^1 - ST_{j_1}^1 \geq \left(\sum_{i \in A_{j_2} \setminus A_{j_1}} s_i^1 - M\left(A_{j_2} \setminus A_{j_1}, n^1 - 1\right)\right) \Big/ n^1 \tag{5a}$$

where n^1 is the number of type 1 providers (nurse practitioners).

Proof. See Appendix A.2.

Proposition 2 provides a lower bound on the time interval between the starting times of any two patients with a nurse practitioner. Since the first patient always sees the nurse practitioner at $t = s_0$, by applying Proposition 2 to the first patient and any other patient j, we can get a lower bound on any patient j's starting time with a nurse practitioner. Adding such constraints to model (1) greatly speeds up the computations because they eliminate many alternative sub-optimal sequences while giving a tighter LP relaxation. These improvements were confirmed during testing.

3.5. A two-step method to solve the deterministic problem

The clinic scheduling problem depicted in Figure 2 is a combination of an open shop and flexible flow shop problem that turns out to be extremely difficult to solve with a commercial code such as CPLEX for more than 10 patients. To obtain solutions, we developed a two-step method that provides both lower and upper bounds as well as a feasible solution to the original problem.

In Step 1, we remove a subset of the original constraints to create a much easier problem. The relaxed solution provides a lower bound on the objective function in (1a) but is rarely, if ever, feasible. In Step 2, we solve a second optimization problem that makes use of the patient sequence found in Step 1. In choosing the constraints to remove in Step 1, we were guided by the speed-up observed after tentatively removing a set of constraints as well as the relative value of the lower bound obtained. For our problem, we found that the best compromise was to remove the following two sets of constraints:

1. Room constraints. There were three reasons for this decision. First, removing the room constraints led to only a small decrease in the objective function value. Second, only minimal violations of the constraints were observed, and third, the problem became much easier to solve since many of the binary variables could also be removed.
2. Nurse practitioner constraints. As the number of providers decreases, the problem gets easier to solve and still provides a lower bound. The decision to omit the nurse practitioners was made for two reasons.

    i. Given that all patients must see a nurse practitioner first, this is the only provider whose waiting time can be taken into account after she/he is removed from the model. In the original problem, the total delay of patient j attributable to a type k provider consists of two parts: (i) service time s_j^k with the provider, and (ii) waiting time when the provider is occupied with prior patients. If we remove the type k provider from the problem without taking into account one or both of these times, the likelihood of getting a strong lower bound is not very high. The advantage of removing the nurse practitioner rather than any of the other providers is that we are able to connect the starting time of a patient's encounter with the nurse practitioner to the patient's arrival time. For example, if patient j arrives at time point $q-2$, and $s_j^1 + adm^1 - 2\tau > 0$, then any patient who sees the same nurse practitioner as j and arrives at time point q would need to wait for at least $s_j^1 + adm^1 - 2\tau$ minutes before seeing this nurse practitioner. This calculation is myopic and therefore provides a lower bound of the true waiting time.

    Let t_j^{wait} denote such a lower bound, and for convenience let $\bar{m} = \max_{j\in J}(s_j^1 + adm^1)/\tau$. The constraints below are needed to determine t_j^{wait}. In each constraint, patient j's waiting time should be no less than the delay caused by prior patients who see the same nurse practitioner:

$$t_j^{wait} \geq \sum_{i\in J} x_{i,q-m}^{arr1} \cdot \left(s_i^1 + adm^1\right) - m\cdot\tau - \left(1 - x_{j,q}^{arr1}\right)\cdot T_{max},$$

$$\forall\ j \in J, m = 1, \ldots, \bar{m} \tag{6a}$$

$$t_j^{wait} \geq \sum_{i\in J} x_{i,q-m}^{arr2} \cdot \left(s_i^1 + adm^1\right) - m\cdot\tau - \left(1 - x_{j,q}^{arr2}\right)\cdot T_{max},$$

$$\forall\ j \in J, m = 1, \ldots, \bar{m} \tag{6b}$$

    ii. Because there are two nurse practitioners and every patient must be seen by one of them, the number of binary variables and constraints needed to model this encounter is much greater than for the other providers. Therefore, removing the nurse practitioners greatly reduces the size of an instance and was seen to reduce runtimes by almost an order of magnitude.

Based on the solution from Step 1, we construct a feasible solution to the original problem in Step 2 by adding back the room constraints and solving a modified optimization problem that makes use of patient order. The details follow.

Two-Step Method

Step 1 **(a)** *Preprocessing*. Modify model (1) as follows: remove all the variables that have index $k = 1$; remove the nurse practitioner constraints, which are those in model (1) for $k = 1$; add constraints (6) to model (1); subtract t_j^{wait} and s_j^k from the right-hand side of constraints (1k) to account for the delay associated with waiting for and being treated by a nurse practitioner after check in.

(b) *Solution*. Set up and solve the relaxed model which consists of the modifications made to model (1) in part (a), constraints (4), and constraints (5a) in Proposition 2.

(c) *Output*. Each patient's appointment time at the clinic. These values can be calculated from x_{jq}^{arr1} and $x_{jq}^{arr2}, \forall\ j \in J,\ \ q = 0, 1, \ldots, T_{max}/\tau - 1$.

Step 2 **(a)** *Preprocessing*. Order the patients based on their arrival time in the solution found in Step 1. Each patient has a rank order.

(b) *Model modifications*. Construct a new model, which includes model (1), constraints (3), constraints (4), and constraints (5a) in Proposition 2. Also add the following constraints: if patient j's rank order is two or more greater than patient i's in the solution found in Step 1, then $\sum_{q=0}^{T_{max}/\tau-1}(x_{jq}^{arr1} + x_{jq}^{arr2})\cdot q\cdot\tau \leq \sum_{q=0}^{T_{max}/\tau-1}(x_{iq}^{arr1} + x_{iq}^{arr2})\cdot q\cdot\tau$. Accordingly, j will arrive no earlier than patient i in the new solution.

(c) *Solution*. Set up and solve the model resulting from the modifications found in part (b).

(d) *Output*. Each patient's appointment time and the schedule template for the clinic.

4. Stochastic model

The deterministic EOSS model formulated in Section 3 can be used to solve an instance of the daily appointment

scheduling problem, but it falls short in accounting for the stochastic elements in the system. Our real goal is to develop an appointment template that is robust in the face of probabilistic service times and patient flows. *A priori* uncertainty in routing is the norm when patients are to be seen by multiple providers in a single visit. In fact, it is the rule rather than the exception in many clinical settings, since the personalized plan of care is made after the patient has been initially interviewed and examined to determine the severity of his/her condition. Therefore, it is not possible to accurately predict which providers he/she will need to see during the visit.

Based on our deterministic model, we have developed a two-stage integer stochastic programming model in which the patient mix along with service times, provider sets and pathways are random variables. The objective of the model is to minimize a weighted combination of expected clinic closing time and patient waiting time over a wide range of scenarios. In the accompanying analysis it is assumed that the no-show rate is zero and that all patients arrive at their scheduled time.

4.1. Stochastic problem

In our IPU scheduling problem, the likelihood that a patient sees a particular provider for a specific amount of time is determined by probability distributions obtained from the Dell Medical School Department of Surgery. For lower extremity joint pain, new and follow-up patients are further divided into six sub-types: (new) mild osteoarthritic, moderate osteoarthritic, severe osteoarthritic, operative, follow-up non-operative, and follow-up operative. Given their proportional mix and their associated probability distributions for provider sets and service times, it is possible to generate scenarios using Monte Carlo sampling. Our original intent was to generate half-day scenarios (4.5-hour clinical sessions) and then try to solve the corresponding two-stage stochastic program to determine the optimal appointment template. We found, however, that as the number of scenarios grew it was increasingly difficult to find solutions, so various alternatives to tackling the full problem were investigated. In the simplest case, we find a template and corresponding patient flow for each scenario separately by solving the corresponding deterministic EOSS model. The average clinic closing time and patient waiting time are then calculated over the different scenarios to get a lower bound on long-run clinic performance. This is called the *wait-and-see* (WS) solution.

In the first stage of the two-stage model, a single appointment template is determined without knowing the patient mix, provider sets, pathways, and service times. In the second stage, this information is revealed for each scenario. To formulate the problem, denote the patients' provider sets and service times by $\tilde{K}$ and $\tilde{s}$, respectively. Assuming for the moment that the appointment template is known, we can then find the optimal arrival times, room occupancy times, and provider start and end times with their patients for each scenario. That is, we can find the optimal values of the second stage variables, which we denote by $\hat{x} \equiv \{x, y, z, x^{arr}, t, ST, T^1, T^2\}$. These values specify each patient's arrival time and schedule with his/her providers. Letting $n^{arr} \equiv (n_0^{arr}, n_1^{arr}, ..., n_{T_{max}/\tau}^{arr})$ be the arrivals at time point q, the two-stage stochastic program, also known as the *recourse problem* (RP), is

$$\min_{n^{arr}} \quad E_{\tilde{s},\tilde{K}}\left[f\left(n^{arr}, \tilde{s}, \tilde{K}\right)\right] \tag{7a}$$

$$\text{s.t.} \quad \text{Constraints (1b)}-\text{(1o)}, \ \text{(3a)}-\text{(3e)} \ \text{and} \ \text{(4a)}-\text{(4d)} \tag{7b}$$

where $E_{\tilde{s},\tilde{K}}$ denotes the expectation with respect to the random variables $\tilde{s}$ and $\tilde{K}$, and $f(n^{arr}, \tilde{s}, \tilde{K})$ is defined as

$$f\left(n^{arr}, \tilde{s}, \tilde{K}\right) = \min_{\hat{x}} \ \alpha_1 \cdot T + \alpha_2 \cdot \sum_{j \in J}\left(T_j^2 - T_j^1\right)$$

The function $f(\cdot)$ represents the second-stage problem. Conceptually, after the appointment template n^{arr} is fixed in the first stage, all uncertainty is resolved and optimal schedules can be determined in the second stage for each patient in each scenario. For a fixed template, the individual scenario instances can be solved separately (we solve each scenario using our deterministic model presented in Section 3) and their objective values averaged to get an approximation of the objective function value in model (7a). This approach is based on sample average approximation (e.g., see Kleywegt *et al.* 2002).

4.2. Solving the stochastic model

When the number of scenarios is finite, the two-stage stochastic program is typically approached by creating a deterministic equivalent one-stage, mixed-integer program. In the reformulation, the second-stage constraints and variables are indexed by scenario and the expected value in model (7a) is replaced with the average of the second-stage objective functions (e.g., see Engell *et al.* 2004; Bard *et al.* 2007). However, such an approach does not always work well because the computational burden increases dramatically as the number of scenarios increases. This was the situation that we faced after enumerating only a few scenarios.

The first alternative that we investigated involved replacing the random parameters with their expected values to obtain a deterministic formulation known at the Expected Value (EV) problem. For IPUs, however, the likelihood of a patient seeing a particular provider follows a probability distribution, so taking the expectation of the patient's provider set would lead to fractional visits. To deal with this situation we conducted a Monte Carlo simulation by sampling each patient's provider set to generate different scenarios. In each scenario, we used the expected service times and expected number of patients of each of the six types (rounded to the nearest integer). The optimization problem for each scenario is solved using our deterministic model in Section 3. After finding the solution for each scenario, we average the numbers of patients who arrive at each time point in all scenarios to get the expected value solution. The EV problem

can be stated as follows:

$$EV = \min_{n^{arr}} \; E_{\tilde{K}}\left[f\left(n^{arr}, \hat{x}, E[\tilde{s}], \tilde{K}\right)\right] \tag{8a}$$

where the optimal objective function value is denoted by EV and the value of the template variables is denoted by n^{arr}_{EV}. We are also interested in the solution of the following three problems which are used to evaluate the quality of the EV solution and to calculate upper and lower bounds on the optimal solution:

$$RP = \min_{n^{arr}} \; E_{\tilde{s},\tilde{K}}\left[f\left(n^{arr}, \hat{x}, \tilde{s}, \tilde{K}\right)\right] \tag{9a}$$

$$EEV = \; E_{\tilde{s},\tilde{K}}\left[f\left(n^{arr}_{EV}, \hat{x}, \tilde{s}, \tilde{K}\right)\right] \tag{9b}$$

$$WS = \; E_{\tilde{s},\tilde{K}}\left[\min_{n^{arr}} f\left(n^{arr}, \hat{x}, \tilde{s}, \tilde{K}\right)\right] \tag{9c}$$

RP represents the two-stage stochastic program given by model (7), and as mentioned, is not solvable; hence the need for bounds. To measure the quality of the EV solution, we fix the template in RP to n^{arr}_{EV} and solve the resulting second-stage problems separately. Averaging their objective function values gives what is called the expected cost of the EV solution, which is denoted by EEV. The EEV value is an upper bound on RP and WS is a lower bound (see Birge and Louveaux (2011)). Thus we have the following relationships:

$$WS \leq RP \leq EEV$$

The optimality gap associated with EV is the gap between EEV and RP, which derives from the loss of stochasticity in the EV problem. The gap between WS and RP results from the loss of perfect information. Neither of these gaps are possible to obtain in our case, though, because we are not able to compute RP for realistic instances. Therefore, we turn to the gap between WS and EEV to evaluate the quality of the EV solution. Since the WS and EEV problems are solved using the two-step method, we use the gap between the step-1 value obtained from the WS problem, and the step-2 value obtained from the EEV problem to evaluate performance.

5. Computational results

All models were implemented in C++ using IBM's Concert Technology library and solved with CPLEX 12.7. The experiments were performed on a Linux workstation with 4 Intel(R) Core(TM) i7-4790 CPU, 8 3.60 GHz processors and 16 GB memory running Ubuntu 16.04. All problem instances discussed in this section were solved optimally using CPLEX's default setting. In constraints (1h) and (1m), the value of ϵ was set to 1.2.

5.1. Data and scenarios

In the analysis, we consider half-day sessions consisting of a fixed number of patients. Arrivals are scheduled by the models at multiples of 15-minute intervals beginning at 8:00 a.m. The total time allocated for check-in and rooming is fixed at 8.3 minutes per patient. The IPU operates with two nurse practitioners and one each of the other provider types. The total number of rooms is seven. Table 1 gives the patient mix and the probability that a particular patient type will be seen by each of the providers. The first encounter for all patients is with a nurse practitioner and the last is with the care planner, both with probability 1, so these providers are omitted from the table. As mentioned, the new patients are divided into four groups and the follow-ups into two groups. The ratio between the new and follow-up patients is 3:1.

Table 1. Patient probabilities for visits with providers.

Patient type	Patient mix	Surgeon	Physical therapist	Nutritionist
New mild osteoarthritis	0.330	0.25	0.5	0.4
New moderate osteoarthritis	0.3225	0.5	0.5	0.4
New severe osteoarthritis	0.056 25	0.9	0.7	0.4
New operative path	0.041 25	1	0.9	0.4
Follow-up non-operative path	0.1875	0.3925	0.4875	0.378
Follow-up operative path	0.0625	1	0.5	0

We model the probabilities for a certain type of patient seeing each of the different providers as independent. This reflects the fact that we do not know a given patient's path *a priori*. Whether a patient sees a certain provider is determined after the patient arrives at the clinic and is examined by the nurse practitioner. Under such circumstances, it is common to take a population-level view and use independently sampled probabilities (White *et al.* 2011; Lahiri and Seidmann 2012; Dobson *et al.* 2013; Saghafian *et al.* 2014). Service time distributions are enumerated in Table 2.

The implied pathways and probability distributions in Tables 1 and 2 are based on estimates provided by the director of the lower extremity joint pain IPU in the musculoskeletal area at the Dell Medical School (DMS) (fourth author on this paper) and other providers from the DMS Department of Surgery who had experience with the same patient population at other clinics prior to the formation of the joint pain IPU. The six patient types (pathways) identified in the two tables represent a common characterization of patients seeking treatment for joint pain. This level of detail allowed the clinical team to estimate the probabilities associated with the resources required to provide care to each type of patient. At the highest level, patients are generally classified as new or follow-up. Clinically speaking, there are only two types of follow-up patients. Those that follow up after surgery and have a certain type of pathway resulting in a fairly short and predictable visit, versus a non-operative follow-up visit, which is similar across disease severity and somewhat longer than a postoperative visit. In rare cases, some patients may benefit from supplementary services such as psychiatry, social work and behavioral health. However, having dedicated providers to cover these services could not be justified financially so they were not included in the design of the IPU.

In the absence of historical data, anecdotal evidence suggests that the time to undergo medical procedures in an outpatient setting can be modeled using minimum, maximum

Table 2. Service time probability distributions (minutes).

Patient type	*Nurse practitioner*	*Surgeon*	*Physical therapist*	*Nutritionist*	*Care planner*
New mild osteoarthritis	Tri(15,20,30)	Tri(7,10,20)	Tri(10,15,25)	Tri(10,15,25)	Tri (5,10,20)
New moderate osteoarthritis	Tri(15,20,30)	Tri(7,10,20)	Tri(10,15,25)	Tri(10,15,25)	Tri (5,10,20)
New severe osteoarthritis	Tri(15,20,30)	Tri(7,10,20)	Tri(10,15,25)	Tri(10,15,25)	Tri (5,10,20)
New operative path	Tri(15,20,30)	Tri(7,10,20)	Tri(10,15,25)	Tri(10,15,25)	Tri (5,10,20)
Follow-up non-operative path	Tri(7,12,17)	Tri(4.3,5.4,10.8)	Tri(10.8,16.2,21.6)	Tri(6,8,12)	Tri (5,10,20)
Follow-up operative path	Tri(7,12,17)	Tri(3.024,4.32,8.64)	Tri(6.48,8.64,12.96)	0	Tri (5,10,20)

Table 3. Optimality gap (%) for the two-step method with 10 patients.

Statistics	*GAP 1*	*GAP 2*	*GAP 2-1*
Mean	2.73	0.97	3.69
HW[1]	0.39	0.21	0.49

[1]Half width of a 95% confidence interval.

Table 4. GAP 2-1 (%) for the two-step method with different numbers of patients.

Number of patients	*7*	*8*	*9*	*10*
Mean	2.38	3.14	3.61	3.69
HW[1]	0.39	0.45	0.44	0.49

[1]Half width of a 95% confidence interval.

and modal times (e.g., see Swisher *et al.* 2001). These three parameters, solicited from the aforementioned providers, lead directly to a triangular distribution, which we use for service times. As an aside, when the clinic opened in the fall of 2017, the staff was able to collect data on provider service times and patient mix. This led to a few adjustments in the probabilistic data in Tables 1 and 2, but for the most part, the original estimates turned out to be highly accurate.

In our evaluation of the two-step method in Section 5.2, parameter values and provider sets for each patient are generated independently. First, we determine which type of patient is being considered by sampling from the patient mix distribution in Table 1. Although the total number of patients is fixed in each scenario, the ratio of new to follow-ups changes from one realization to the next. After each patient's group is determined, we generate his/her provider set based on the probabilities in Table 1, and service times from the triangular distributions in Table 2. The same generated data sets are used for the EEV and WS problems.

When deriving the EV template defined in Section 5.3.1, rather than sampling from the patient mix distribution, the numbers of new and follow-up patients were set to their approximate expected values. For each patient type, the provider set was generated based on the probabilities in Table 1, while the expected service time with each provider was taken as the weighted sum (the weight is the patient mix fraction) of the mean service time. For example, the expected service time of a follow-up patient with the surgeon is the weighted sum of the mean of the bottom two triangle distributions under the column "Surgeon" in Table 2. Lastly, our models reflect whether a patient is new or making a follow-up appointment at the time of booking. In practice, this is all the information that is available to the scheduling clerk.

5.2. Two-step method

In the first set of experiments, our goal was to evaluate the quality of the solutions obtained with the two-step method presented in Section 3.5 for solving the deterministic model. We began by randomly generating 200 instances (scenarios) with 10 patients each and then applying the algorithm. The number of patients in each instance was determined by sampling from a multinomial distribution with probabilities {0.3, 0.2, 0.1, 0.1, 0.2, 0.1}, which approximates the patient mix in Table 1. Similarly, the provider set for each type of patient was sampled using the probabilities in Table 1 while the service times were sampled from the triangular distributions in Table 2. Recall that Step 1 provides a lower bound and Step 2 provides an upper bound on the objective function in (1a). Performance was measured by the percentage deviation from the optimum obtained by solving model (1) as modified to represent the joint pain IPU. We only considered instances with 10 patients in this part of the analysis, as it was not possible to reliably solve larger instances with CPLEX. Note that after 200 instances, the output statistics discussed below were unchanged to two decimal places, indicating that there was no further need for additional sampling. In all, 16 096 seconds were required to find the exact optima for the 200 instances compared with 1935 seconds when using the two-step method to find the bounds.

For each scenario, we calculated the gap between the Step 2 objective function value and the Step 1 value (GAP 2-1), the gap between the Step 2 value and optimal value (GAP 2), and the gap between the optimal value and the Step 1 value (GAP 1). The differences were then converted to percentages and averaged over the 200 scenarios. The results are summarized in Table 3.

From the table we see that the average gap between the bounds found in Steps 1 and 2 is 3.69%, an indication of the strength of the heuristic. Additional evidence of its strength can be seen by examining the percent difference between the upper bound and the optimal solution (GAP 2), which is only 0.97% on average. Moreover, the optimal solution is much closer to the Step 2 solution than the Step 1 solutions because GAP 2 is a third the size of GAP 1. Taken together, these results support the use of the two-step method to derive appointment schedules under more realistic scenarios.

To check the sensitivity of the performance of the two-step method, we repeated the above process for cases with seven, eight and nine patients. The results are reported in Table 4. The optimality gap decreased slightly as the number of patients decreased but remained stable. In our testing with 14 patients in the remaining sections, the gap was always less than 5%.

5.3. Finding robust templates

Our primary goal is to derive a single appointment template whose implementation will ensure clinic durations of less

than 4.5 hours and patient visit times not exceeding 1.5 hours, on average. The recourse problem was designed to achieve this goal, but the computational difficulties we encountered when trying to solve it led to our reliance on the two-step method. The best we can do with this heuristic, however, is to solve a deterministic version of model (1). The approach we take to circumvent this limitation is described below. For the remaining analysis, we work with 14 patients, which is the number that the joint pain IPU would like to schedule each half-day session.

5.3.1. Generating EV templates

Ordinarily, only a single EV template exists, which would be derived by replacing all random parameters in the IPU model with their expected values and then solving problem (7). This was not possible for our problem because the expected number of providers who see a patient is a random variable whose expected value is fractional. As mentioned, Monte Carlo sampling was used to skirt this issue. The first step was to generate a representative number of scenarios by using the data in Table 1 to obtain the provider set for each patient. As an integral approximation to the patient mix, we assumed that each scenario consisted of 11 new patients and three follow-up patients. For the former group, the number of patients of each type was fixed at four, four, two and one. For the latter group, the number of patients was fixed at two and one. We then used the two-step method to find feasible schedules and their corresponding templates n^{arr}, where n^{arr} is a vector that specifies the number of patients who arrive at each 15-minute time point.

To derive a single appointment template, we began by averaging the number of patients who arrive at each time point over all scenarios. Again we found that the output statistics became stable after 200 scenarios, so we terminated the generation process at that point. The total time required to solve the 200 instances was 114 minutes. Figure 3 depicts the results after averaging. The horizontal axis indicates the time points and the vertical axis identifies the average number of patients who are scheduled to arrive at the start of each 15-minute interval.

Rounding strategies. As expected, heights of the bars in the figure are fractional in the figure are fractional, but to be implementable the number of patients must be zero, one or two at each time point, as in the individual solutions. To achieve integrality, a rounding strategy is necessary. The approach we take is based on the observation that the number of patients who arrive earlier in the session affect the statistics of patients who arrive later. Accordingly, the procedure we adopt is to round fractions (up or down), fix the number of patients at one point at a time starting at zero, and sequentially moving forward in 15-minute increments until closing time is reached. At each time point t, the number of patients who have arrived previously is fixed by rounding. We then round the fractional number at t and repeat the procedure at $t+1$.

In particular, after fixing the number of patients who arrive at t, we re-solve the reduced EV problem with the remaining patients, average the results from the newly derived 200 templates, and then round the value at $t+1$.

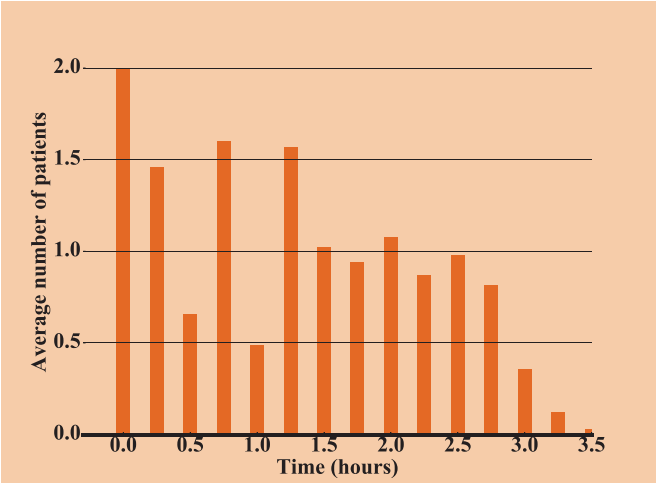

Figure 3. Average number of patients scheduled to arrive at each time point.

For example, at $t=0$, we see in Figure 3 that the average number of patients is very close to two so we fix the number of patients who are scheduled to arrive at $t=0$ to two; that is, we set $n_0^{arr}=2$. We then re-solve the EV problem and take the average of the 200 templates just derived. The corresponding figure is almost identical to Figure 3, so with $n_0^{arr}=2$, we fix n_1^{arr} to be either one or two, depending on the rounding strategy (to follow). After fixing n_1^{arr}, we re-solve the EV problem and move on to n_2^{arr}, and so on.

The number of possible templates increases exponentially with the number of time points for arbitrary rounding. We considered two strategies to generate two templates. In the first strategy we always round up at t unless the fraction is zero or within a small range of an integer value. Based on empirical testing, we chose the cutoff to be 0.2. If the average number of patients is less than 0.2, we round it to zero; if it is between 0.2 and 1.2, we round it to one; if it exceeds 1.2 but is less than two, we round it to two. Without a cutoff we found that the resulting schedules were too aggressive, in that they emphasized earlier appointment times, which led to significantly longer patient waiting times.

In the second strategy, we always round down at each time point, unless the fraction is within the cutoff range. Based on empirical testing, we again chose the cutoff to be 0.2. If the average number of patients is less than 0.8, we round it to zero; if it is between 0.8 and 1.8, we round it to one; if it exceeds 1.8 but is less than two, we round it to two.

The template produced by the first strategy is more aggressive than the second, but rounding down does not always avoid long waits and extended clinic hours. Figure 4 shows the less aggressive EV template (Figure 4(a)) and the more aggressive EV template (Figure 4(b)). Each figure indicates the number of patients scheduled to arrive at each time point. Note that during construction, the last patient in the less aggressive template actually arrives at $t=3.25$. For practical reasons, though, we modified the template slightly to avoid a gap at $t=3.0$ and to conform with what is called the 2BEG schedule in the literature (Cayirli and Veral 2003). Testing showed negligible differences between results produced by the less aggressive template and 2BEG.

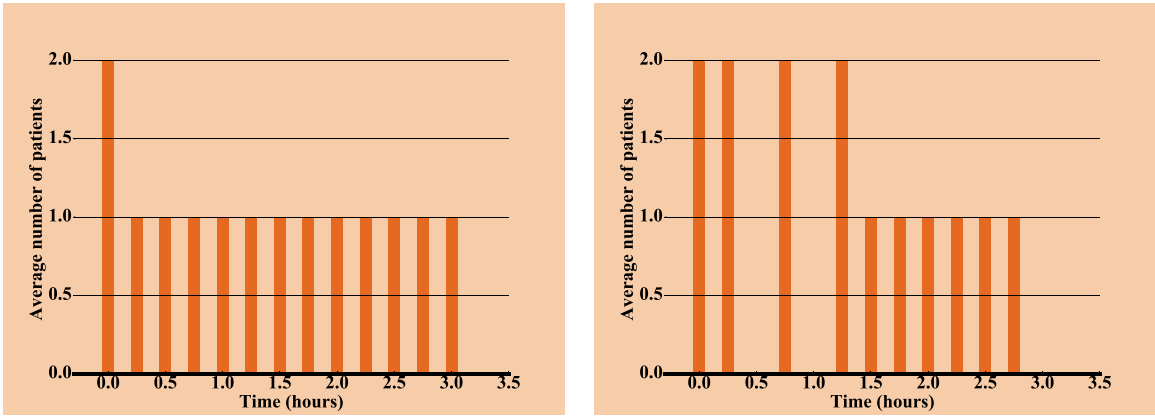

Figure 4. Templates derived from the EV solution: (a) less aggressive EV template and (b) more aggressive EV template.

Comparison of strategies. To visualize the difference between the more aggressive and less aggressive templates, we generated the cumulative number of patients who arrive at the clinic up to each time point t. Of course, the total number of arrivals for the less aggressive template is no greater than that for the more aggressive template at any t. Figure 5 plots the results as a function of time for both templates. Any other template that is constructed from a combination of the less aggressive and more aggressive strategies would be bounded by these two curves. Comparing the cumulative number of patients for the two schedules at any time t shows that the difference is small. In fact, the two plots in Figure 5 indicate that the difference at any time t is either zero or one.

Additional templates. In addition to the two templates derived above, we also evaluated a third from the literature and a fourth based on a variation of the more aggressive template in Figure 4(b). Each of the four templates is formally defined below and consists of the number of patients who arrive between $t = 0$ and $t = 3$ (i.e., between 8 am and 11 am), followed by its name and description.

1) 2-1-1-1-1-1-1-1-1-1-1-1-1: 2BEG. Assigns two patients at the beginning of the session and then one at each point thereafter. It was first proposed and studied in Bailey (1952), and turns out to be the less aggressive EV template that we derived.
2) 2-0-2-0-2-0-2-0-2-0-2-0-2: VBFI-1. VBFI stands for 'variable block/fixed interval,' which means that a different number of patients can be assigned at each time point as long as they are separated by the same fixed interval (see Wijewickrama (2006)). Here, two patients are scheduled to arrive every half hour. This is less aggressive than 2BEG, which can be transformed into VBFI-1 by moving one patient at every other time point to the next time point starting at $t = 0.25$. In our experience, VBFI-1 is commonly used in practice.
3) 2-2-0-2-0-2-1-1-1-1-1-1-0: EV-RU. This is the more aggressive template shown in Figure 4(b), where RU stands for "round up."

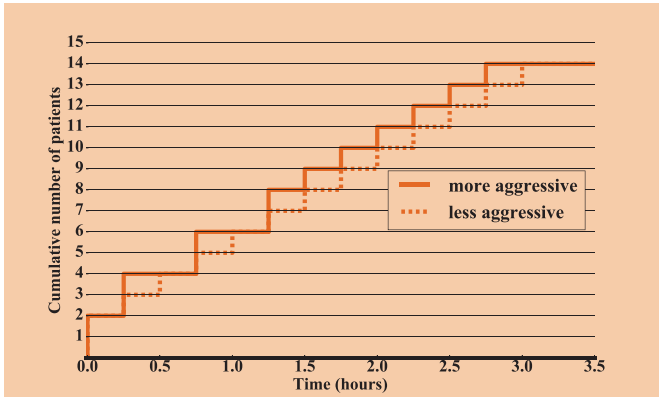

Figure 5. Cumulative number of patient arrivals over a half-day session.

4) 2-2-0-2-0-2-2-0-2-0-2-0-0: VBFI-2. This template is based on EV-RU, but is more aggressive. If we move the patients at $t = 1.75$, 2.25, and 2.75 in the EV-RU template one interval earlier, we can get VBFI-2. Including this template in the study will tell us whether a significant improvement results by making the EV-RU template more aggressive.

5.3.2. Results for candidate templates

To compare the quality of the solutions resulting from the use of each of the four templates, we randomly generated additional scenarios by sampling from the distributions in Tables 1 and 2 to obtain provider sets and service times, respectively, for each patient. The output statistics became stable after 800 scenarios so we stopped at that point. The number of new and follow-up patients were also sampled, although the total number was fixed at 14. To gauge performance, we averaged the objective function values and other metrics over all 800 scenarios for each template. The results are highlighted in Tables 5 and 6 along with the results for the WS problem using the same data. The columns in the tables are arranged from the least aggressive to the most aggressive template. For each template, computation times for all 800 scenarios ranged from a total of 8 to 19 hours.

Table 5. Results for different appointment templates.

	WS		VBFI-1		2BEG		EV-RU		VBFI-2	
Metrics	*Mean*[1]	*HW*[2]	*Mean*	*HW*	*Mean*	*HW*	*Mean*	*HW*	*Mean*	*HW*
Step 1 closing time	4.097	0.028	4.438	0.021	4.343	0.022	4.226	0.023	4.205	0.024
Step 2 closing time	4.295	0.024	4.444	0.020	4.360	0.022	4.289	0.023	4.289	0.023
Feasible rate	800/800		800/800		800/800		794/800		790/800	
Waiting time	0.296	0.006	0.281	0.008	0.300	0.009	0.371	0.009	0.414	0.009
Service time	1.080	0.005	1.080	0.005	1.080	0.005	1.079	0.005	1.079	0.005
Time in clinic	1.376	0.010	1.361	0.012	1.381	0.012	1.450	0.013	1.493	0.013
Waiting time (new)	0.310	0.007	0.284	0.009	0.307	0.009	0.381	0.010	0.424	0.010
Service time (new)	1.162	0.005	1.162	0.005	1.162	0.005	1.161	0.006	1.161	0.006
Time in clinic (new)	1.472	0.011	1.446	0.012	1.469	0.013	1.542	0.013	1.585	0.013
Waiting time (follow-up)	0.242	0.010	0.258	0.010	0.268	0.011	0.319	0.012	0.360	0.012
Service time (follow-up)	0.779	0.007	0.779	0.007	0.779	0.007	0.778	0.007	0.779	0.007
Time in clinic (follow-up)	1.020	0.013	1.037	0.013	1.047	0.014	1.097	0.014	1.138	0.014
Fraction above closing time	NA	NA	0.351	0.033	0.263	0.031	0.217	0.029	0.219	0.029

[1]All times in hours; the statistics are all Step 2 results except for the Step 1 closing time.
[2]Half width of a 95% confidence interval.

Table 6. Resource utilization.

	WS		VBFI-1		2BEG		EV-RU		VBFI-2	
Metrics	*Mean*	*HW*	*Mean*	*HW*	*Mean*	*HW*	*Mean*	*HW*	*Mean*	*HW*
Nurse practitioner 1	0.731	0.004	0.704	0.004	0.724	0.004	0.732	0.004	0.726	0.004
Nurse practitioner 2	0.712	0.004	0.688	0.004	0.696	0.005	0.712	0.004	0.717	0.004
Surgeon	0.485	0.009	0.468	0.008	0.478	0.008	0.486	0.009	0.486	0.009
Physical therapist	0.686	0.010	0.663	0.010	0.676	0.010	0.684	0.010	0.683	0.010
Nutritionist	0.444	0.011	0.429	0.010	0.437	0.011	0.445	0.011	0.445	0.011
Care planner	0.686	0.005	0.662	0.004	0.675	0.004	0.687	0.005	0.687	0.005
Room	0.549	0.003	0.541	0.003	0.548	0.003	0.569	0.003	0.575	0.003

The first two rows in Table 5 report the Step 1 and Step 2 closing times. The remaining rows give the Step 2 flow time statistics for all patients, and then for new patients and follow-ups separately. The last row reports the fraction of cases in which the closing time exceeded 4.5 hours. Table 6 shows provider and room utilizations. As the model includes constraints (1m), which restrict the total time a patient can spend in the clinic to a given maximum, a handful of instances turned out to be infeasible. For the EV-RU template, 6 out of 800 were infeasible and for the VBFI-2 template, 10 out of 800 were infeasible.

Theoretically, the Step 2 closing time obtained from VBFI-2 should be no later than the closing time provided by EV-RU for two reasons: (i) VBFI-2 is more aggressive than EV-RU; and (ii) more infeasible cases are discarded when VBFI-2 is used, which should bring down the average closing time. This follows because late clinic closing times are a result of long patient waiting times, which produce infeasible instances. Nevertheless, the two templates have virtually identical Step 2 closing times, so neither reason was seen to have a noticeable impact on clinic performance. This suggests that the EV-RU template is sufficiently aggressive and that moving to the more aggressive VBFI-2 template will not provide any benefit. This also suggests that there is no bias in the results after discarding the infeasible cases.

Clinic closing time. The first observation from the statistics in Table 5 is that the difference between the Step 1 and Step 2 closing times is less than 2% for all four templates. Although the Step 2 closing time in each case is not necessarily optimal, given that the two-step method was used for the computations, the size of the gap indicates that it should be a very good approximation. One way to evaluate the four

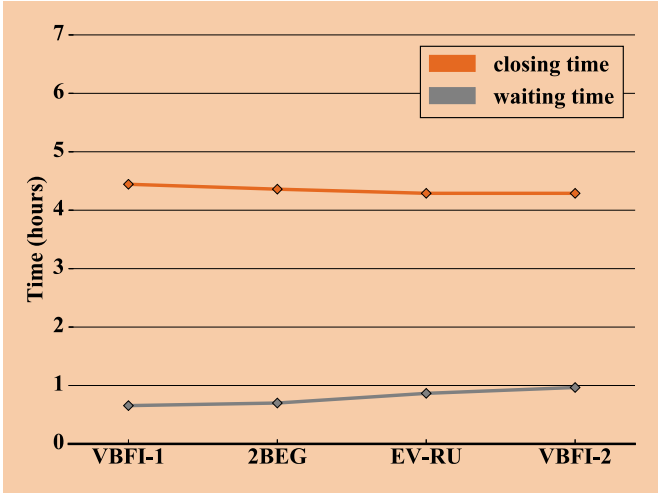

Figure 6. Comparison of four arrival templates.

sets of results is to compare the mean and half width of a 95% confidence interval of clinic closing time of the Step 2 solution. For example, the Step 2 results imply that the 2BEG 95% confidence interval extends from 4.338 to 4.382, whereas the range of the average clinic closing time for EV-RU is from 4.266 to 4.312. As the two confidence intervals do not overlap, we can conclude that the average closing time obtained from the EV-RU template is significantly smaller than the value associated with the 2BEG template.

Another way to compare the closing time for different templates is to check the Step 1 and Step 2 solutions. For example, the lower bound on closing time for 2BEG obtained at Step 1 is 4.343, which is greater than the Step 2 closing time of EV-RU. As such, the true value of closing

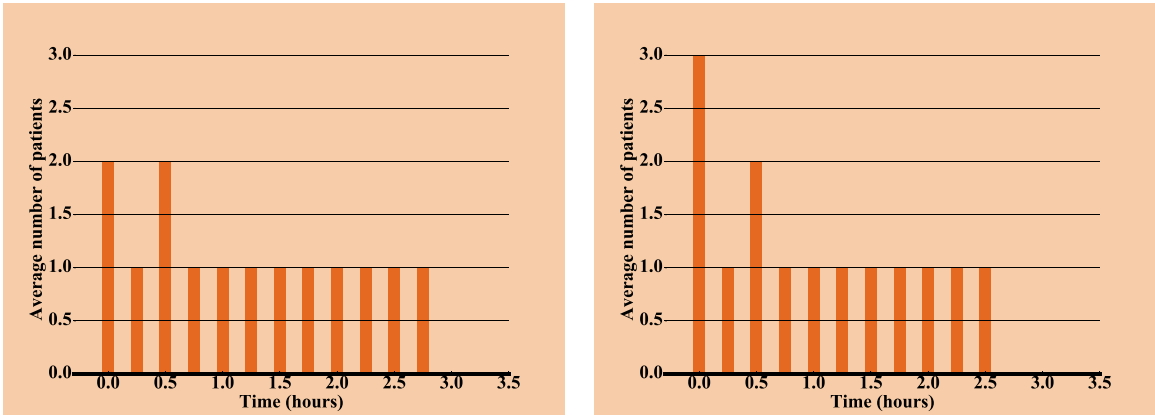

Figure 7. Two templates for the case with three nurse practitioners: (a) VBFI-3: less aggressive EV template and (b) VBFI-4: more aggressive EV template.

time for 2BEG should also be greater than the true value of closing time for EV-RU. By implication, using the EV-RU template should yield lower clinic closing times than the 2BEG template. For the WS problem, its optimal clinic closing time should be no greater than the closing time obtained from any template. As can be seen in Table 5, however, the average WS Step 2 closing time is 4.295, which is greater than 4.289, the average closing time obtained from the EV-RU and VBFI-2 templates. This result is possible because the two-step method only provides feasible solutions. As it turns out, many of the WS solutions are suboptimal.

A second observation about the statistics in Table 5 is that as the templates get more aggressive, the clinic closing times decrease; see Figure 6. This follows because patients generally arrive earlier when the more aggressive templates are used, and are seen earlier by their providers. Hence, they are more likely to finish their visit sooner. As the same 800 scenarios were used in all the computations, the service times are the same across all templates, so the comparative closing time results should not be affected by those values. The statistics in Table 5 confirm that the average service time for a visit is nearly identical for all templates, as well as for the WS problem.

Waiting times and time in clinic. The average waiting time and average total time in the clinic increase as the template becomes more aggressive. Again, more patients arriving earlier makes it more likely that they will face longer queues in front of their providers. This is true for all patients taken as a whole, for new patients, as well as for follow-ups. For example, the waiting time increases from 0.281 to 0.3 to 0.371 to 0.414 hours as the template becomes more aggressive. As might be expected, the average waiting time for the WS problem is relatively small, even though its average closing time is also small. This follows because a separate template is derived for each scenario allowing patient arrivals to better match provider availability.

As the appointment template becomes more aggressive, the waiting time and closing time move in opposite directions, as can be seen in Figure 6. Based on their relative importance, the clinic director can choose the template that achieves the best balance. For example, if preference is given to the closing time, then the EV-RU template may be a good candidate because its closing time is 4.289, which is measurably less than the corresponding value of 4.36 for 2BEG and 4.444 for VBFI-1, a 1.6% (4.3 minutes) and 3.5% (9.3 minutes) reduction, respectively. Moreover, patient waiting times resulting from the EV-RU template increase by 4.3 minutes and 5.4 minutes over 2BEG and VBFI-1, respectively.

The EV-RU template appears to be a good compromise with respect to the primary metrics. If we make it more aggressive by transforming it into the VBFI-2 template, the clinic closing time remains about the same, but the patient waiting times increase significantly. Nevertheless, if the waiting time is relatively more important than the closing time, then the 2BEG template may be a good choice, as its average waiting time of 0.3 hours is somewhat less than the corresponding values of 0.371 for the EV-RU and 0.414 for VBFI-2 templates. In practice, it is not desirable to choose a template less aggressive than 2BEG such as VBFI-1. The reduction in waiting time provided by the latter is only 1.14 minutes on average, whereas the average jump in closing time is 5.04 minutes.

Fraction above target closing time. From Table 5 we see that the fraction of scenarios in which the clinic closing time exceeds the target of 4.5 hours decreases for the first three templates as they get more aggressive. For VBFI-1 the percentage is 35.1, whereas for EV-RU the percentage drops to 21.7. There is almost no difference between EV-RU and VBFI-2, which gives further evidence that VBFI-2 does not improve clinic performance even though it is more aggressive than EV-RU.

Utilization. Table 6 reports the utilization for the six individual providers and the seven rooms. Although there are some statistically significant differences between the templates for each provider type, they are negligible in practice. The contrast in room utilization is a bit sharper, but still negligible. Note that the values in the table are based on the time the first patient arrives and the last patient leaves. At first glance, the statistics may be somewhat misleading, as it takes over an hour for the clinic to fill up and roughly the same amount of time for it to empty out. While waiting times average up to 25 minutes, for example, room utilization is less than 60% on average. This supposed

contradiction, can be explained by the transient effects at the beginning and end of the session.

5.3.3. Different resource levels

To determine the potential value of increasing or decreasing resource levels, we investigated two possibilities. In particular, nurse practitioners and rooms are two resources that afford some leeway in clinic design. Preliminary testing suggested that decreasing or increasing the number of rooms by one barely affected system performance, whereas increasing the number of nurse practitioners by one had a noticeable impact. Consequently, in this section we only present results for three nurse practitioners.

In the analysis, we followed the same procedure outlined in Section 5.3.1 using the same data for the patient mix and service time distributions. The two templates shown in Figure 7 parallel those in Figure 4. The VBFI-3 template is the less aggressive of the two and VBFI-4 is the more aggressive. Our previous results for these templates still hold. For example, the VBFI-3 template provides better outcomes if the patient waiting time has more weight than the clinic closing time, and the VBFI-4 template is better if clinic closing time is the more important metric.

It is more interesting, though, to compare the system with two and three nurse practitioners. The statistical results for these new templates are highlighted in Tables 7 and 8. A comparison with the statistics in Tables 5 and 7 indicates that adding one nurse practitioner significantly reduces both the clinic closing time and the patient waiting time. For the more aggressive templates, for example, the clinic closing time decreases from 4.289 to 4.005 hours (6.6%), and the patient waiting time decreases from 0.371 to 0.351 hours (5.4%). Nevertheless, whether the financial investment required to achieve this performance boost can be justified, is still an open question.

With respect to resource utilization, the nurse practitioners are the bottleneck when two are present because they have the highest utilization among all providers. When a third one is added, the bottleneck switches to the physical therapist and the care planner whose utilizations are now over 70%. In light of these statistics, adding a fourth nurse practitioner cannot be justified.

5.3.4. Appointment rules

For the joint pain IPU, follow-up patients represent roughly 25% of the flow. In several recent studies, it has been shown that ordering the patients in the schedule by type can improve clinic performance (e.g., see Bosch and Dietz 2000; White *et al.* 2011). In this section, we propose several rules that derive from our observations of arrival patterns associated with each template for the original case with two nurse practitioners. Figure 8 contains four graphs that plot the average number of patients in each of the two groups who arrive at the beginning of each 15-minute interval. The graphs were constructed using the same data set that provided the computational results in Table 5. In this part of the analysis, our objective is to gain insight into how the model chooses appointment slots for new compared with follow-up patients under the various templates.

Table 7. Results for different appointment templates.

	VBFI-3		*VBFI-4*	
Metrics	*Mean* [1]	*HW*[2]	*Mean*	*HW*
Step 1 closing time	4.147	0.023	3.990	0.026
Step 2 closing time	4.149	0.023	4.005	0.025
Feasible rate	800/800		800/800	
Waiting time	0.291	0.009	0.351	0.010
Service time	1.080	0.005	1.080	0.005
Time in clinic	1.371	0.013	1.432	0.014
Waiting time (new)	0.307	0.010	0.375	0.011
Service time (new)	1.162	0.005	1.162	0.005
Time in clinic (new)	1.469	0.014	1.536	0.015
Waiting time (follow-up)	0.216	0.012	0.249	0.014
Service time (follow-up)	0.779	0.007	0.779	0.007
Time in clinic (follow-up)	0.995	0.014	1.027	0.016
Fraction above closing time	0.134	0.024	0.086	0.020

[1]All times in hours; the statistics are all Step 2 results except for the Step 1 closing time.
[2]Half width of a 95% confidence interval.

Table 8. Resource utilization.

	VBFI-3		*VBFI-4*	
Metrics	*Mean*	*HW*	*Mean*	*HW*
Nurse practitioner 1	0.570	0.005	0.582	0.005
Nurse practitioner 2	0.508	0.005	0.528	0.005
Nurse practitioner 3	0.416	0.005	0.440	0.005
Surgeon	0.502	0.009	0.521	0.009
Physical therapist	0.709	0.010	0.734	0.010
Nutritionist	0.460	0.011	0.477	0.012
Care planner	0.710	0.005	0.736	0.005
Room	0.578	0.003	0.620	0.003

Since the ratio of follow-up to new patients is 3:11, statistically, the expected number of follow-ups at each time point is the total number of patients multiplied by 3/14. By comparing the average number of follow-ups at each time point with the expected number, we can find the time slots when they have a high chance of being scheduled to arrive. For example, the expected number of follow-up patients at $t = 0$ for EV-RU is $2 \cdot 3/14 = 3/7$. In our experiments, we found that the average number of follow-ups that arrive at $t = 0$ for EV-RU is around 0.6, which is greater than 3/7. Therefore, we say the follow-up patient has a higher chance of being scheduled to arrive at $t = 0$ for EV-RU than might be expected. Similar analysis can be done for the other time points and templates.

The following patterns appear in the graphs in Figure 8.

Pattern 1: A follow-up patient has a high chance of arriving at the beginning of the session.

There are explanations for this pattern: (i) follow-up patients usually have shorter service times than new patients. Starting with one new patient and one follow-up will generally result in the latter finishing the nurse practitioner visit sooner and then moving on to her next provider. Thus, the next provider will be engaged sooner than if both patients at $t = 0$ were from the same group. Moreover, when the new patient finishes his/her visit with the nurse practitioner, if he/she is required to see the same provider as the follow-up, then his wait will likely be shorter; (ii) the difference in expected service times between the first two patients creates

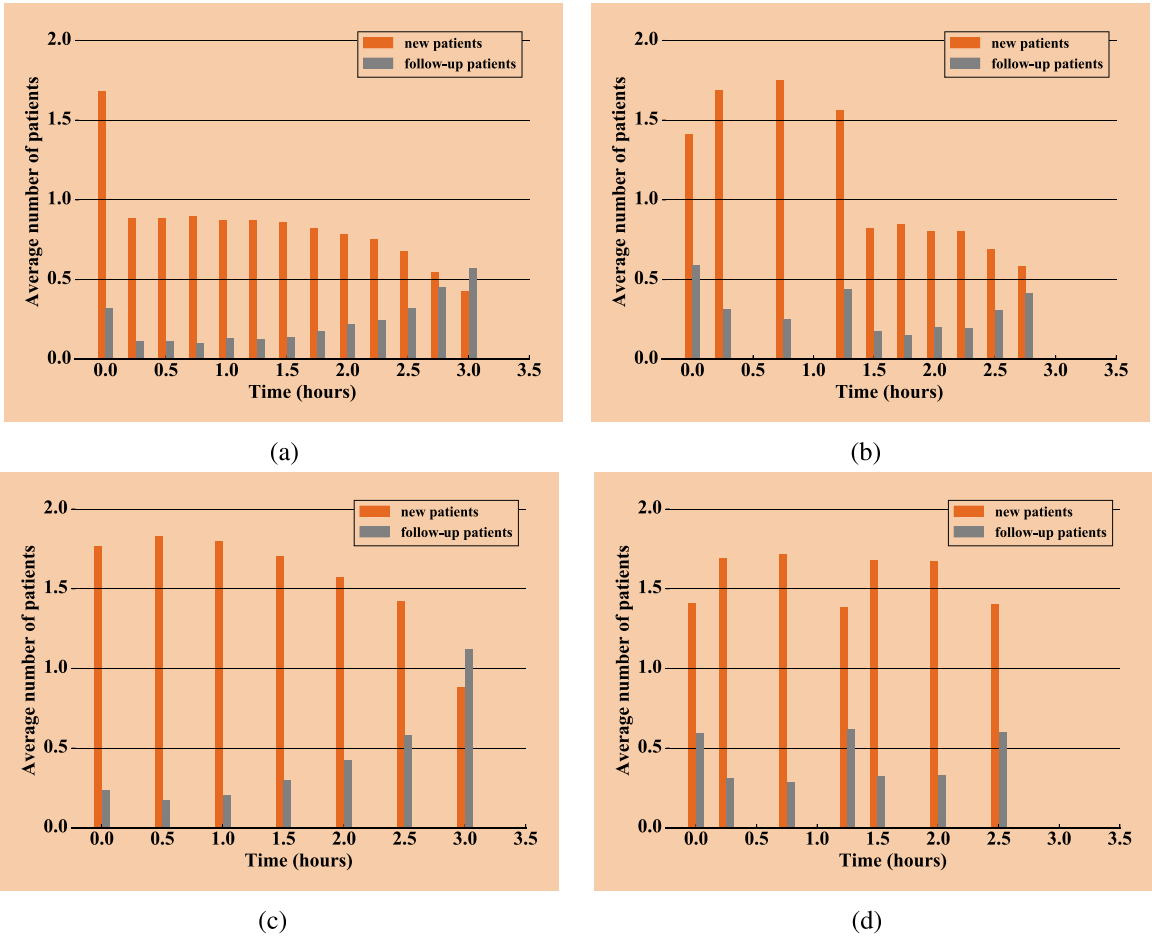

Figure 8. Average number of new and follow-up patients for different templates for the case with two nurse practitioners: (a) 2BEG; (b) EV-RU; (c) VBFI-1; and (d) VBFI-2.

a staggered flow with respect to downstream providers. This tends to reduce congestion as well as the clinic closing time.

Rule 1: Schedule both a follow-up patient and a new patient at $t = 0$.

Pattern 2: When there are three or more patients scheduled at two successive time points, one of them is a follow-up patient.

At most time points, only a single patient is scheduled to arrive. At some time points in some templates, though, the patient flow can be high. In template EV-RU, for example, the total number of new and follow-up patients who arrive at successive time points $t = 1.25$ and $t = 1.5$ is three; for VBFI-2, the total number who arrive at $t = 1.25$ and $t = 1.5$ is four. In such cases, congestion is likely leading to long queues in front of the providers. By scheduling a follow-up patient to arrive at those time points with high inflow, the likelihood of congestion will be reduced because follow-ups typically spend less time with providers.

The second reason to schedule a follow-up patient to arrive at time points where the patient inflow is more than two is that all rooms are likely to be occupied. Again, follow-up patients usually spend less time with providers, and so will spend less time in the clinic. This will help limit queuing for rooms.

Rule 2: Embedded in the statement of Pattern 2.

Pattern 3: A follow-up patient has a high chance of arriving at the end of the session.

Pattern 3 appears in the results for all four templates. This can be explained as follows. Assume that there are 13 patients in the system and queues exist for all providers other than the two nurse practitioners. Consider the extreme case where the 14th arrival is a new patient who is to be seen by all five providers. In this scenario, it is likely that the care planner has already finished his/her consultation with the first 13 patients before the 14th patient finishes with his/her fourth provider. The idle time between the 13th and 14th patients has the effect of delaying the clinic closing time. If a follow-up patient is the last to arrive, however, it is less likely that the care planner will have finished consulting with the previous 13 patients, as service times for follow-up patients are less than for new patients.

Rule 3: The last appointment should be a follow-up patient.

To check the robustness of the above patterns, an additional set of experiments was conducted to determine whether they still hold for the case with three nurse practitioners. The results are depicted in Figure 9. Indeed, Patterns 1 and 3 are still present in Figure 9, but Pattern 2 has disappeared. The absence of Pattern 2 is a consequence of increased capacity due to the additional nurse

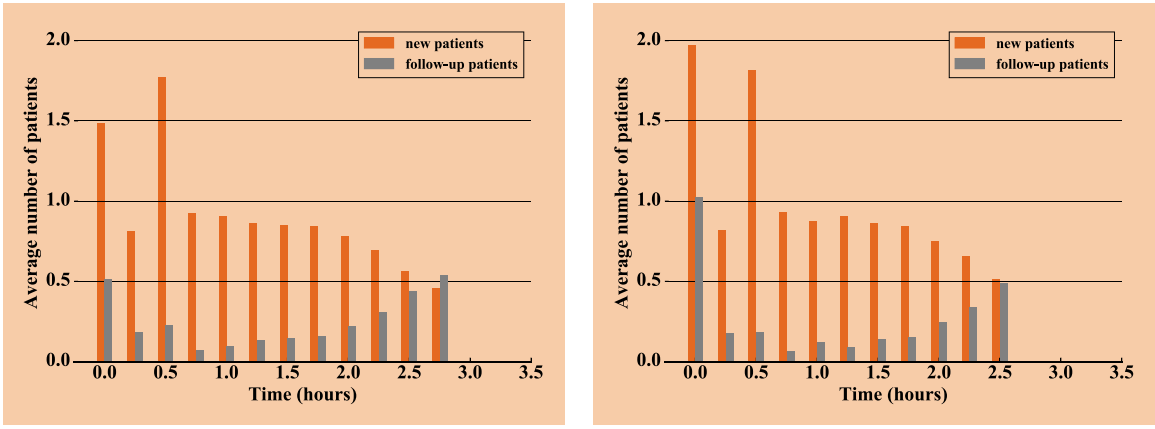

Figure 9. Average number of new and follow-up patients for different templates for three nurse practitioners: (a) VBFI-3 and (b) VBFI-4.

practitioner. Therefore, even when three patients are scheduled to arrive at two successive time points, there will be little if any queueing in front of any of the nurse practitioners. Hence, there is no need to schedule a follow-up patient at either time point to improve flow.

Of course, it may not be possible in practice to fully adhere to these rules due to requests for specific appointment times, provider availability, or the random nature of the patient mix. However, they do provide some level of insight and guidance for improving clinic efficiency. In our experience, outpatient scheduling is typically done on a first-come, first-served basis without taking into account patient type.

6. Summary and conclusions

The complexity of patient flow in multi-provider clinics such as IPUs underscores the need for a considered approach to appointment scheduling to maximize the use of available resources while ensuring high levels of customer satisfaction. In this article, we first proposed a new model for the extended open shop problem, and then tailored it to an IPU in which multiple patient and provider types have to be coordinated over the day. For the deterministic version of the problem, we developed a two-step method that provides solutions for 10 patients within 4% of optimality on average. These results were derived by analyzing a wide-range of scenarios reflecting operations of the joint pain IPU at the DMS. A two-stage integer stochastic optimization model was then presented that more realistically represents actual patient-provider interactions. The two-step method was again used to solve the WS problem and several versions of the EV problem. All instances contained 14 patients. The average optimality gap was less than 5% for the WS problem and less than 2% for the EV variants. Our ultimate goal has been to determine an appointment template that can be used to schedule new and follow-up patients over half-day sessions.

The results from our experimental design indicated that the templates derived from the proposed methodology provide good performance with respect to minimizing a combination of clinic closing time and patient waiting time. The relatively less aggressive templates (i) VBFI-1 (variable block/fixed interval), which allows a different number of patients to be assigned at each time point as long as they are separated by the same fixed interval, and (ii) 2BEG, where two patients are scheduled at the beginning of the session and then a single patient at fixed intervals thereafter, are preferable if patient waiting time is the clinic's primary metric. The more aggressive template EV-RU (expected value-rounded up) is more effective when the clinic closing time is of primary importance. We also observed arrival patterns by patient type for each template, and proposed several scheduling rules based on the insights gained. For example, one follow-up and one new patient should be scheduled to arrive at the beginning of the day, and one follow-up at the end. In general, similar patterns were observed in two of the three cases when we increased the number of nurse practitioners from two to three. Collectively, these results have provided the foundation for designing the DMS joint pain IPU schedule.

One limitation of our model is that it does not account for the stochasticity of the arrival process. When patients depart from their scheduled appointment times by arriving early or late, the result is more uncertainty, which can lead to increased system congestion, longer queues and sojourn times, and later closing times. The greater the uncertainty, the greater the disruption to the planned schedule. A second limitation of our work is that we did not consider patient no shows. Due to the need to coordinate multiple providers and patient types in an IPU, any disruptions in the flow can create measurable inefficiencies in clinic operations. When we began our study, we did not have the necessary data to postulate no-show probabilities for any of the six patient types, as we were designing a new clinic. Rather than guessing we decided to assume that all patients arrive for their appointment on time. This allowed us to design templates for the ideal case. Further investigation and data collection are needed to determine the most effective way of dealing with no shows. Existing approaches typically resort to overbooking or shortening appointment slots to reduce the negative consequences of absent patients. However, there is no

standard way of implementing either of these ideas that reliably minimizes the disruption to the system.

Notes on contributors

Pengfei Zhang is a fourth year Ph.D. student in the McCombs School of Business at the University of Texas at Austin. He graduated from the School of the Gifted Young at the University of Science and Technology of China with a bachelor's degree in physics. He obtained his master's degree in industrial engineering from the University of Arizona. His research interests include modeling of healthcare delivery systems and robust optimization.

Jonathan F. Bard is a professor of operations research & industrial engineering in the Mechanical Engineering Department at the University of Texas at Austin. He holds the Industrial Properties Corporation Endowed Faculty Fellowship, and serves as the Associate Director of the Center for the Management of Operations and Logistics. He received a D.Sc. in Operations Research from The George Washington University. Dr. Bard's research centers on improving healthcare delivery, personnel scheduling, production planning and control, and the design of decomposition algorithms for solving large-scale optimization problems, and has appeared in a wide variety of technical Journals. Currently, he serves on six editorial boards and previously was a Focused Issue Editor of IIE Transactions and an Associate Editor of Management Science. He is a registered engineer in the State of Texas, a fellow of IIE and INFORMS, and a senior member of IEEE. In the past, he has held a number of offices in each of these organizations, and is currently the INFORMS Vice President of Publications.

Douglas J. Morrice holds the Bobbie and Coulter R. Sublett Centennial Professorship in Business. He is also Professor of Operations Management and a University of Texas Supply Chain Management Center of Excellence Senior Research Fellow in the McCombs School of Business at The University of Texas at Austin. Dr. Morrice has an ORIE Ph.D. from Cornell University. His research interests include simulation design, modeling, and analysis, healthcare delivery management, and supply chain risk management. He is a senior editor for Production and Operations Management, an editor-at-large for Interfaces, and an area editor for IISE Transactions on Healthcare Systems Engineering.

Karl M. Koenig is the Medical Director of the Musculoskeletal Institute at Dell Medical School and leads the Integrated Practice Unit for Joint Pain. He also leads the effort to develop IPUs for other musculoskeletal conditions including Back Pain, Fracture Care, Sports Medicine, and Foot Care. After receiving his undergraduate degree at the Massachusetts Institute of Technology, he attended Baylor College of Medicine. Dr. Koenig completed his residency at Dartmouth-Hitchcock Medical Center (Lebanon, NH) in Orthopaedic Surgery and fellowship in Adult Reconstruction at Stanford University. He is also a graduate of the Dartmouth Institute for Health Policy and Clinical Practice, where he began his work on patient outcomes and value-based healthcare delivery. Prior to joining Dell Medical School, Dr. Koenig led the Division of Adult Reconstruction at Dartmouth for 5 years and was one of the architects of GreenCare, a sweeping quality improvement initiative to create a self-improving microsystem around total joint replacement.

Funding

This project was supported by the Dell Medical School's Texas Health Catalyst program. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the University of Texas at Austin. Additional support was provided by a grant from the McCombs School of Business at the University of Texas.

References

Ahmadi-Javid, A., Jalali, Z. and Klassen, K.J. (2017) Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research*, **258**(1), 3–34.

Azadeh, A., Baghersad, M., Farahani, M.H. and Zarrin, M. (2015) Semi-online patient scheduling in pathology laboratories. *Artificial Intelligence in Medicine*, **64**(3), 217–226.

Bailey, N.T. (1952) A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. *Journal of the Royal Statistical Society. Series B (Methodological)*, **14**(2), 185–199.

Bard, J.F., Morton, D.P. and Wang, Y.M. (2007) Workforce planning at UPSS mail processing and distribution centers using stochastic optimization. *Annals of Operations Research*, **155**(1), 51–78.

Bard, J.F., Shu, Z., Morrice, D.J., Wang, D., Poursani, R. and Leykum, L. (2016) Improving patient flow at a family health clinic. *Health Care Management Science*, **19**(2), 170–191.

Berg, B.P., Denton, B.T., Erdogan, S.A., Rohleder, T. and Huschka, T. (2014) Optimal booking and scheduling in outpatient procedure centers. *Computers & Operations Research*, **50**, 24–37.

Bhat, P.B., Prasanna, V.K. and Raghavendra, C.S. (2000) Block-cyclic redistribution over heterogeneous networks. *Cluster Computing*, **3**(1), 25–34.

Birge, J.R. and Louveaux, F. (2011) *Introduction to Stochastic Programming*, Springer Science & Business Media, Springer-Verlag, New York, NY.

Bosch, P.M.V. and Dietz, D.C. (2000) Minimizing expected waiting in a medical appointment system. *IIE Transactions*, **32**(9), 841–848.

Castro, E. and Petrovic, S. (2012) Combined mathematical programming and heuristics for a radiotherapy pre-treatment scheduling problem. *Journal of Scheduling*, **15**(3), 333–346.

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

Cayirli, T., Veral, E. and Rosen, H. (2006) Designing appointment scheduling systems for ambulatory care services. *Health Care Management Science*, **9**(1):47–58.

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

Chen, R.R. and Robinson, L.W. (2014) Sequencing and scheduling appointments with potential call-in patients. *Production and Operations Management*, **23**(9), 1522–1538.

CMS. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsHistorical.html. Accessed 18 July 2018.

Denton, B.T., Miller, A.J., Balasubramanian, H.J. and Huschka, T.R. (2010) Optimal allocation of surgery blocks to operating rooms under uncertainty. *Operations Research*, **58**(4, Part 1 of 2), 802–816.

Dobson, G., Tezcan, T. and Tilson, V. (2013) Optimal workflow decisions for investigators in systems with interruptions. *Management Science*, **59**(5), 1125–1141.

Engell, S., Märkert, A., Sand, G. and Schultz, R. (2004) Aggregated scheduling of a multiproduct batch plant by two-stage stochastic integer programming. *Optimization and Engineering*, **5**(3), 335–359.

Erdogan, S.A. and Denton, B. (2013) Dynamic appointment scheduling of a stochastic server with uncertain demand. *INFORMS Journal on Computing*, **25**(1), 116–132.

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

Gupta, D. and Wang, L. (2008) Revenue management for a primary-care clinic in the presence of patient choice. *Operations Research*, **56**(3), 576–592.

Hanna, A. (2010) Patient-centred care. *Ontario Medical Review*, **1**, 34–49.

Kazemian, P., Sir, M.Y., van Oyen, M.P., Lovely, J.K., Larson, D.W. and Pasupathy, K. S. (2017) Coordinating clinic and surgery appointments to meet access service levels for elective surgery. *Journal of Biomedical Informatics*, **66**, 105–115.

Keswani, A., Koenig, K.M. and Bozic, K.J. (2016) Value-based healthcare: Part 1–designing and implementing integrated practice units

for the management of musculoskeletal disease. *Clinical Orthopaedics and Related Research*, **474**(10), 2100–2103.
Kleywegt, A.J., Shapiro, A. and Homem-de Mello, T. (2002) The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, **12**(2), 479–502.
Kong, Q., Lee, C.-Y., Teo, C.-P. and Zheng, Z. (2013) Scheduling arrivals to a stochastic service delivery system using copositive cones. *Operations Research*, **61**(3), 711–726.
Lahiri, A. and Seidmann, A. (2012) Information hang-overs in healthcare service systems. *Manufacturing & Service Operations Management*, **14**(4), 634–653.
Leeftink, A.G., Bikker, I.A., Vliegen, I.M.H. and Boucherie, R.J. (2018) Multi-disciplinary planning in health care: A review. *Health Systems*. Advance online publication. doi:10.1080/20476965.2018.1436909
Liaw, C.-F. (2000) A hybrid genetic algorithm for the open shop scheduling problem. *European Journal of Operational Research*, **124**(1), 28–42.
Mancilla, C. and Storer, R. (2012) A sample average approximation approach to stochastic appointment sequencing and scheduling. *IIE Transactions*, **44**(8), 655–670.
Mannino, C., Nilssen, E.J. and Nordlander, T.E. (2012) A pattern based, robust approach to cyclic master surgery scheduling. *Journal of Scheduling*, **15**(5), 553–563.
Miller, C.E., Tucker, A.W. and Zemlin, R.A. (1960) Integer programming formulation of traveling salesman problems. *Journal of the ACM*, **7**(4), 326–329.
Noori-Darvish, S., Mahdavi, I. and Mahdavi-Amiri, N. (2012) A bi-objective possibilistic programming model for open shop scheduling problems with sequence-dependent setup times, fuzzy processing times, and fuzzy due dates. *Applied Soft Computing*, **12**(4), 1399–1416.
Oh, H.-J., Muriel, A., Balasubramanian, H., Atkinson, K. and Ptaszkiewicz, T. (2013) Guidelines for scheduling in primary care under different patient types and stochastic nurse and provider service times. *IIE Transactions on Healthcare Systems Engineering*, **3**(4), 263–279.
Parizi, M.S. and Ghate, A. (2016) Multi-class, multi-resource advance scheduling with no-shows, cancellations and overbooking. *Computers & Operations Research*, **67**, 90–101.
Pérez, E., Ntaimo, L., Malavé, C.O., Bailey, C. and McCormack, P. (2013) Stochastic online appointment scheduling of multi-step sequential procedures in nuclear medicine. *Health Care Management Science*, **16**(4), 281–299.
Pinedo, M.L. (2016) *Scheduling: Theory, Algorithms, and Systems*. Springer, New York, NY.
Porter, M.E. (2010) What is value in health care? *New England Journal of Medicine*, **363**(26), 2477–2481.
Qu, X., Peng, Y., Kong, N. and Shi, J. (2013) A two-phase approach to scheduling multi-category outpatient appointments–a case study of a women's clinic. *Health Care Management Science*, **16**(3), 197–216.
Rachuba, S. and Werners, B. (2014) A robust approach for scheduling in hospitals using multiple objectives. *Journal of the Operational Research Society*, **65**(4), 546–556.
Saghafian, S., Hopp, W.J., van Oyen, M.P., Desmond, J.S. and Kronick, S.L. (2014) Complexity-augmented triage: A tool for improving patient safety and operational efficiency. *Manufacturing & Service Operations Management*, **16**(3), 329–345.
Stewart, M., Brown, J.B., Donner, A., McWhinney, I.R., Oates, J., Weston, W.W. and Jordan, J. (2000) The impact of patient-centered care on outcomes. *Family Practice*, **49**, 796–804.
Swisher, J.R., Jacobson, S.H., Jun, J.B. and Balci, O. (2001) Modeling and analyzing a physician clinic environment using discrete-event (visual) simulation. *Computers & Operations Research*, **28**(2), 105–125.
Truong, V.-A. (2015) Optimal advance scheduling. *Management Science*, **61**(7), 1584–1597.
Wang, D., Morrice, D.J., Muthuraman, K., Bard, J.F., Leykum, L.K. and Noorily, S.H. (2018) Coordinated scheduling for a multi-server network in outpatient pre-operative care. *Production and Operations Management*, **27**(3), 458–479.
White, D.L., Froehle, C.M. and Klassen, K.J. (2011) The effect of integrated scheduling and capacity policies on clinical efficiency. *Production and Operations Management*, **20**(3), 442–455.
Wijewickrama, A.K.A. (2006) Simulation analysis for reducing queues in mixed-patients' outpatient department. *International Journal of Simulation Modelling*, **5**(2), 56–68.

Appendices

A. Proof of Proposition 1 and Proposition 2

A.1. Proof for Proposition 1

We first show the validity of constraints (2a) when $n=1$. Suppose that patient j's position in provider k's schedule is m_j+1. Then $x^k_{j,m_j+1}=1$ and $ST^k_j = t^k_{m_j+1}$, and by implication, $x^k_{j,m+1}=0$ for $m=0,1,\ldots,m_j-1,m_j+1,\ldots,m_k-1$. Three cases are possible for the constraints (2a).

(1) $m=0,1,\ldots,m_j-1$. Here, $x^k_{j,m+1}=0$ so the first summation on the Right-Hand Side (RHS) of constraints (2a) becomes zero and the second becomes $T_{max}-LT^k_{m^k-m_j+m}$. As a consequence, constraints (2a) reduce to

$$t^k_{m_j+1} = ST^k_j \le t^k_{m+1} + T_{max} - LT^k_{m^k-m_j+m}$$

Note that there are m_j-m patient encounters for provider k between t^k_{m+1} and $t^k_{m_j+1}$, and that the remaining $m^k-(m_j-m)$ encounters start during the time period $(t^k_{m+1}-0)+(T_{max}-t^k_{m_j+1})$. By definition, $LT^k_{m^k-m_j+m}$ equals the sum of the m^k-m_j+m smallest service times of provider k's patients, which means that it is a lower bound on the total time for any combination of $m^k-(m_j-m)$ encounters. Accordingly, $LT^k_{m^k-m_j+m} \le (t^k_{m+1}-0)+(T_{max}-t^k_{m_j+1})$, which validates the above inequality.

(2) $m=m_j$. Here, $x^k_{j,m+1}=1$ so both the first and second summation on the RHS of constraints (2a) become zero. As a consequence, constraints (2a) reduces to the following inequality given that $t^k_{m_j+1}=t^k_{m+1}$:

$$t^k_{m_j+1} = ST^k_j \le t^k_{m+1}$$

(3) $m=m_j+1,\ldots,m^k-1$. Here, $x^k_{j,m+1}=0$ so the first summation on the RHS of constraints (2a) becomes $LT^k_{m-m_j}$ and the second becomes 0. Thus, constraints (2a) reduce to the following:

$$t^k_{m_j+1} = ST^k_j \le t^k_{m+1} - LT^k_{m-m_j}$$

For provider k, there are $m-m_j$ patient encounters starting between $t^k_{m_j+1}$ and t^k_{m+1}. Again by definition, $LT^k_{m-m_j}$ equals the sum of the $m-m_j$ smallest service times of provider k's patients and is a lower bound on the total time for any $m-m_j$ encounters. Therefore, we have $LT^k_{m-m_j} \le t^k_{m+1}-t^k_{m_j+1}$, which validates the above inequality.

Next we prove that constraints (2a) are actually stronger than their counterparts in constraints (1i). For constraints (2a) and any value of m between 0 and m^k-1, we have:

$$ST^k_j \le t^k_{m+1} + \sum_{m'\le m-1}\left(-LT^k_{m-m'}\right)\cdot x^k_{j,m'+1} + \sum_{m'\ge m+1}\left(T_{max}-LT^k_{m^k-m'+m}\right)\cdot x^k_{j,m'+1} \quad \text{(A1a)}$$

$$\le t^k_{m+1} + \sum_{m'\le m-1} T_{max}\cdot x^k_{j,m'+1} + \sum_{m'\ge m+1} T_{max}\cdot x^k_{j,m'+1} \quad \text{(A1b)}$$

$$= t^k_{m\cdot n^k+1} + \left(\sum_{m'\ne m} x^k_{j,m\cdot n^k+1}\right)\cdot T_{max} \quad \text{(A1c)}$$

$$= t^k_{m\cdot n^k+1} + \left(1-x^k_{j,m\cdot n^k+1}\right)\cdot T_{max}, \quad m=0,\ldots,m^k-1, n^k=1, j\in J, \forall\ k\in K(j) \quad \text{(A1d)}$$

The proofs for the remaining inequalities are identical. □

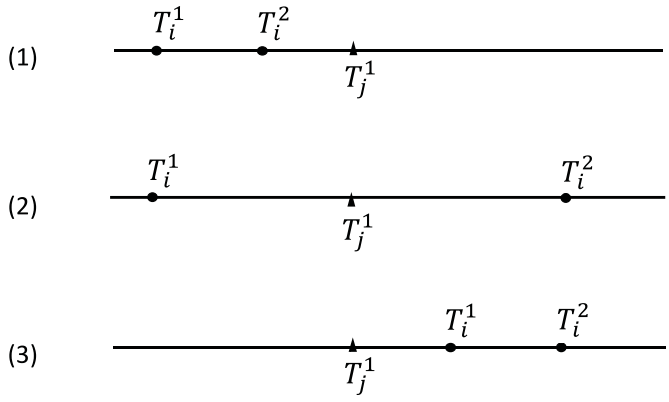

Figure A1. An example for entering-checking method.

A.2. Proof of Proposition 2

We only provide a proof for the case in which there are two nurse practitioners. The arguments for the general case are similar. For the case with $n^1 = 2$ nurse practitioners, we need to show that if patient j_1 starts no later than patient j_2, then $ST^1_{j_2} - ST^1_{j_1} \geq (\sum_{i \in A_{j_2} \setminus A_{j_1}} s^1_i - \max_{i \in A_{j_2} \setminus A_{j_1}} s^1_i)/2$.

Suppose patients $i_1, i_2, \ldots, i_m \in A_{j_2} \setminus (A_{j_1} \cup \{j_1\})$. All patients whose visit with a nurse practitioner is no earlier than $ST^1_{j_1}$ and no later than $ST^1_{j_2}$ are $j_1, i_1, i_2, \ldots, i_m, j_2$. Of these patients, there is at most one whose encounter with a nurse practitioner ends no earlier than $ST^1_{j_2}$ besides patient j_2. In other words, all patients $j_1, i_1, i_2, \ldots, i_m$ start no earlier than ST_{j_1}, and at most one them finishes no earlier than $ST^1_{j_2}$. Suppose that patient i^*, finishes no earlier than $ST^1_{j_2}$. Then,

$$2 \cdot \left(ST^1_{j_2} - ST^1_{j_1}\right) \geq s^1_{j_1} + s^1_{i_1} + s^1_{i_2} + \ldots + s^1_{i_m} - s^1_{i^*}$$
$$\geq s^1_{j_1} + s^1_{i_1} + s^1_{i_2} + \ldots + s^1_{i_m} - \max\left\{s^1_{j_1}, s^1_{i_1}, s^1_{i_2}, \ldots, s^1_{i_m}\right\}$$

or

$$ST^1_{j_2} - ST^1_{j_1} \geq \left(\sum_{i \in A_{j_2} \setminus A_{j_1}} s^1_i - \max_{i \in A_{j_2} \setminus A_{j_1}} s^1_i\right) \Big/ 2$$

□

B. Room constraints

Additional decision variables

- η_{ij} 1 if $T^1_i \leq T^1_j$, which means patient i is placed in a room no later than patient j is placed in a room (the rooms for i and j can be different), 0 otherwise
- η'_{ij} 1 if $T^2_i \leq T^1_j$, which means patient i finishes using a room no later than patient j starts to use a room (the rooms for i and j can be different), 0 otherwise
- ζ_{ij} 1 if patients i and j use the same room, and patient j follows (not necessarily immediately) patient i, 0 otherwise
- δ^r_j 1 if patient j uses room r, 0 otherwise

B.1. Entering-checking method

This method ensures that there is a room available for patient j when service starts with his/her first provider. For any other patient i who has previously entered the clinic, we already know his/her starting time T^1_i and ending time T^2_i, so we already know the values of η_{ij} and η'_{ij}. This information allows us to determine the number of occupied rooms when the patient j sees his first provider. To ensure that a room is available, this number must be less than the total number of rooms, R.

Proposition B.1. *A necessary and sufficient condition that arriving patient j can be placed in a room is that*

$$\sum_{i \neq j, i \in J} \left(\eta_{ij} - \eta'_{ij}\right) \leq R-1, \forall j \in J.$$

Proof. For any patient j whose visit starts with his first provider at time T^1_j, we need to show that the above inequality is satisfied if a room is available. That is, we need to determine how many of the R rooms are occupied. Now, for any other patient i, the three cases shown in Figure A1 need to be considered:

1. $T^1_i < T^1_j, T^2_i \leq T^1_j$. In this case, we have $\eta_{ij} = 1$ and $\eta'_{ij} = 1$, so the room used by patient i is available for patient j.
2. $T^1_i \leq T^1_j, T^2_i > T^1_j$. In this case, we have $\eta_{ij} = 1$ and $\eta'_{ij} = 0$, indicating that patient i is still in his room so it is not available for patient j.
3. $T^1_i > T^1_j, T^2_i > T^1_j$. In this case, we have $\eta_{ij} = 0$ and $\eta'_{ij} = 0$, implying that patient i has not yet been placed in a room, so whichever room he/she is eventually assigned is immaterial to a room being available for patient j. Of course, the time in clinic for patients i and j may overlap, which implies that they cannot use the same room. This will be ensured when a check is made for patient i to determine if a room is available, but it is not a concern when patient j is being assigned a room.

From these cases, we see that when patient j's encounter with his/her first provider begins, if $\eta_{ij} - \eta'_{ij} = 1$, then patient i is using a room; if $\eta_{ij} - \eta'_{ij} = 0$, then patient i is not using a room. Accordingly, when patient j enters the clinic at time T^1_j, the total number of rooms that are being used is $\sum_{i \neq j, i \in J}(\eta_{ij} - \eta'_{ij})$. If patient j can be placed in a room, then the total number of rooms that are being used must be no more than $R - 1$. In contrast, if the total number of rooms that are being used is R, then patient j cannot be placed in a room. Therefore, $\sum_{i \neq j, i \in J}(\eta_{ij} - \eta'_{ij}) \leq R-1$ is a necessary and sufficient condition that a room is available for patient j. □

Based on Proposition B.1, we have the following constraints for the room requirement.

$$T^1_i \geq T^1_j - \eta_{ij} T_{max}, \forall\ i \neq j \in J \tag{A2a}$$

$$T^1_j \geq T^1_i - (1-\eta_{ij}) \cdot T_{max}, \forall\ i \neq j \in J \tag{A2b}$$

$$T^2_i \geq T^1_j - \eta'_{ij} T_{max}, \forall\ i \neq j \in J \tag{A2c}$$

$$T^1_j \geq T^2_i - \left(1-\eta'_{ij}\right) \cdot T_{max}, \forall\ i \neq j \in J \tag{A2d}$$

$$\sum_{i \neq j, i \in J} \left(\eta_{ij} - \eta'_{ij}\right) \leq R-1, \forall\ j \in J \tag{A2e}$$

$$\eta_{ij} + \eta_{ji} \geq 1, \forall\ i \neq j \in J \tag{A2f}$$

$$\eta_{ij} \geq \eta'_{ij}, \forall\ i \neq j \in J \tag{A2g}$$

$$\sum_{m=1}^{m^k} m \cdot x^k_{jm} \geq \sum_{m=1}^{m^k} m \cdot x^k_{im} + 1 - m^k \cdot \left(1-\eta'_{ij}\right), \forall\ i,j \in J(k),$$
$$\forall\ k \in \{k : n^k = 1\} \tag{A2h}$$

$$\eta_{ij}, \eta'_{ij} \in \{0,1\}, \forall\ i,j \in J \tag{A2i}$$

Constraints (A2a) and (A2b) ensure that $\eta_{ij} = 1$ when patient i is placed in a room no later than patient j, and 0 otherwise. Constraints (A2c) and (A2d) ensure $\eta'_{ij} = 1$ if patient i finishes using his/her room before patient j is placed in a room, and 0 otherwise. Constraints (A2e) guarantee that the total number of rooms being used when

patient j is placed in a room is no greater than R – 1. Constraints (A2f) specify that either patient i starts no later than j, or patient j starts no later than i. This is needed for the case in which patients i and j are placed in different rooms at the same time. Without (A2f), η_{ij}, η'_{ij}, η_{ji}, and η'_{ji} will all be zero when rooming occurs simultaneously for the two patients.

Constraints (A2g) are useful cuts which impose the restriction that if patient i finishes earlier than patient j starts, then patient i must also start earlier than patient j. Constraints (A2h) are also useful cuts, which state that if patient i finishes earlier than patient j starts, then for any provider who is the only provider of his/her type, patient i's position index should be smaller than patient j's position index. The difference must be at least one. Constraints (A2i) define the variables as binary.

B.2. Not-immediate-successor method

We begin by assigning each patient to a room. For any two patients who are assigned to a same room, we use binary variables to ensure that they do not overlap in time. That is, if two patients are assigned to the same room, then the starting time of the successor (not necessarily immediate successor) can be no earlier than the ending time of all his predecessors:

$$\sum_{r=1}^{R} \delta_j^r = 1, \forall\ j \in J \tag{A3a}$$

$$\zeta_{ij} + \zeta_{ji} \geq \delta_i^r + \delta_j^r - 1, \forall\ i \neq j \in J, r = 1, \ldots, R \tag{A3b}$$

$$T_j^1 \geq T_i^2 - \left(1 - \zeta_{ij}\right) \cdot T_{max}, \forall\ i \neq j \in J \tag{A3c}$$

$$\sum_{m=1}^{m^k} m \cdot x_{jm}^k \geq \sum_{m=1}^{m^k} m \cdot x_{im}^k + 1 + \sum_{m \in J(k), m \neq i, m \neq j} \left(\zeta_{im} + \zeta_{mj} - 1\right) - \left(1 - \zeta_{ij}\right) \cdot m^k,$$
$$\forall\ i \neq j \in J(k), \forall\ k \in \{k \in K : n^k = 1\} \tag{A3d}$$

$$\zeta_{ij} \in \{0, 1\}, \forall\ i, j \in J \tag{A3e}$$

Constraints (A3a) ensure that each patient has a room. Constraints (A3b) specify that two patients who are assigned to the same room must use the room in sequence. Constraints (A3c) enforce the requirement that the starting time of patient j cannot be earlier than the ending time of all his/her predecessors i who are assigned the same room. Constraints (A3d) and (A3e) parallel (3d) and (3e).