



# Appointment scheduling in multi-stage outpatient clinics

Kenneth J. Klassen<sup>1</sup> · Reena Yoogalingam<sup>2</sup>

Received: 16 May 2017 / Accepted: 23 January 2018 / Published online: 5 February 2018  
© Springer Science+Business Media, LLC, part of Springer Nature 2018

## Abstract

Healthcare providers can benefit from adding less costly capacity to their existing resources in order to satisfy demand while maintaining the quality of patient care. The addition of mid-level service providers (MLSPs) such as physician assistants or nurse practitioners that carry out portions of patient care provides a viable alternative for adding physician capacity. This research considers the circumstances under which adding an MLSP to a single-physician outpatient office becomes the best strategy for the clinic, and determines how scheduling policies from the widely-researched single-stage environment should be adjusted for a multi-stage environment. Compared to a single-stage system where a physician completes all portions of the service, we show that adding an MLSP can reduce patient waiting time, patient flow time, and physician service time with patients. This, in turn, can enable the clinic to see more patients and/or free up physician time for other tasks. Appointment scheduling rules are developed for a multi-stage outpatient service system using a simulation optimization approach. Performance measures focus on the patient experience and clinic operation before and during *each* stage of service.

**Keywords** Appointment scheduling · Simulation optimization · Multi-stage health systems

## 1 Introduction

Healthcare systems are facing increasing costs, a larger number of users, a population that is more conscious about healthcare issues, and increasing demand for quality care [7]. In addition, physicians are in short supply in many areas, a situation which is predicted to deteriorate over time [16]. Consequently, health care providers are looking for ways to add less costly capacity to their existing resources in order to satisfy demand and reduce waiting times while maintaining the quality of patient care. Inefficiency in outpatient clinics in terms of long patient waiting time, idle time, and overtime of the clinic typically results from a mismatch between demand for physician services and capacity available. This problem is further complicated by uncertainty in interarrival times of

patients and service times. Developing an effective appointment schedule can contribute to a better patient experience in terms of reduced waiting time and improved flow through the system. Clinics often attempt to develop schedules that reduce the amount of pre-service waiting time which many patients find frustrating [12]. This may have the added benefit of improving patient perceptions of the quality of service provided [35].

In a single-physician environment, one strategy for improving medical clinic operations is to introduce an additional stage to the service environment where a secondary service provider carries out some of the patient care. This is becoming more common as clinics seek to reduce costs by employing medical professionals that have a lower cost than physicians. The addition of mid-level service providers (MLSPs) such as physician assistants and nurse practitioners is a viable alternative for adding capacity [26]. This strategy can improve patient flow time, enable the clinic to see more patients, and/or free up some of the physician's time for other tasks.

This study considers an appointment scheduling problem in a multi-stage outpatient clinic where a MLSP carries out some of the patient care before the physician sees the patient. With multiple stages, the focus necessarily has to be on the patient experience before and during *each* stage of service. Both factors will determine the patient's level of satisfaction with the service provided. In other words, introducing an

---

✉ Reena Yoogalingam  
ryoogalingam@brocku.ca

Kenneth J. Klassen  
kklassen@brocku.ca

<sup>1</sup> Goodman School of Business, Brock University, 500 Glenridge Ave, St Catharines, ON L2S 3A1, Canada

<sup>2</sup> Goodman School of Business, Brock University 500 Glenridge Ave, St Catharines, ON L2S 3A1, Canada

MLSP to reduce the pre-service wait only to have the patient wait an excessive amount of time to see a physician would be counterproductive. Coordination is more complex since the performance measures for subsequent stages are now dependent on performance at earlier stages of service. Service and arrival times in this environment are stochastic at each stage, which further complicates the design of an efficient system [30]. This research considers the circumstances under which adding an MLSP to a single-physician office is the best strategy for a clinic, develops the best schedule and sequencing of patients in these cases, and determines how policies from the widely-researched single-stage environment should be adjusted for a multi-stage environment. Findings could apply to a variety of multi-stage service systems, such as a dental assistant and dentist, legal assistant and lawyer, or other environments where a trained assistant can complete a subset of required tasks before the more highly trained individual performs more critical tasks. For example, this could reflect a situation where a dental assistant performs tasks such as preparation, cleanings, and x-rays which allows more time for the dentist to perform tasks involving fillings, extractions, and crowns.

The term MLSP is defined more broadly than what is understood as a “mid-level service provider” [35]. In a health care context it includes those trained as a registered nurse, nurse practitioner, or physician assistant. It is assumed that the MLSP has training to carry out some aspects of diagnosis, treatment, and referrals (i.e., more than just take height, weight and blood pressure).

Previous studies have used simulation or analytical methods to address appointment scheduling problems. Analytical studies have typically focused on finding an optimal scheduling rule [9, 17, 27]. These studies are often limited in terms of the complexity of the systems they can account for including the number of stages they consider and distributions (e.g., Exponential, Erlang) used to represent the stochastic parameters. Simulation studies are able to capture greater complexity of the systems [5, 6, 15, 18]. However, they have more difficulty finding an optimal policy for the system.

More recently, several studies have integrated simulation with an optimization technique to address single stage appointment scheduling problems. Many of these studies combine simulation with a metaheuristic technique (e.g., [21, 22]). This approach is well suited to appointment scheduling problems since good solutions can be found while simultaneously incorporating the significant sources of uncertainty that are present. In particular, a simulation optimization approach allows for the consideration of more complex distributions for interarrival and service times and arrival processes at each stage of service in an outpatient clinic. However, with the exception of [35], where simulation optimization is used to solve a multi-stage operating room scheduling problem, there have been relatively few studies that have considered its use in

a multi-stage pre-scheduled environment. In this study, a simulation optimization approach is used to determine how a variety of factors including clinic size, patient unpunctuality, cost structure, allocation strategies of patients to MLSPs, and corresponding reductions in physician service times impact performance. General scheduling rules specifying the length of appointment intervals as well as the sequence of appointments are developed taking these factors into account for a multi-stage outpatient clinic. Data from a multi-stage outpatient clinic is used as the basis for testing a wide range of parameters.

This paper is organized as follows. In the next section a review of the related literature is provided. Then, an overview of the data collected for this study and the problem formulation are discussed. Finally, results and analysis are presented, ending with a discussion and managerial implications.

## 2 Related literature

Appointment scheduling in health care facilities has been widely studied. The vast majority of studies have focused on a single stage appointment scheduling environment, comprising a body of research that spans more than 60 years [4, 13]. These studies primarily focus on several key measures to evaluate the performance of the appointment system. The objective is typically to minimize a weighted sum of patient waiting times, clinic idle time, and/or overtime for a single stage of the process [5, 6, 31]. Each of these studies tested different combinations of measures such as waiting time plus day end time [40, 41], customer waiting time plus physician idle time [32], and waiting time, idle time, and overtime [9, 34]. Other measures such as patient flow times [24, 40, 41] and physician utilization [38] have also been studied.

The goal of this body of work has been to improve performance of health care systems through the design of appointment scheduling “rules”. These rules specify the length of each appointment interval and the number of patients to schedule in that interval (block size). Many innovative rules have been developed with different combinations of fixed and variable appointment interval lengths and block sizes [4]. These include fixed interval rules such as Bailey’s rule which specifies fixed intervals with two patients in the first slot and subsequent patients scheduled in individual blocks [2]. A number of recent studies have focused on optimizing appointment schedules with variable-interval rules. The most prominent finding has been the “dome” rule consisting of individual blocks with initially short appointment intervals, gradually increasing toward the middle of the session, and decreasing toward the end of the session [9, 31, 41, 42]. In [20], the authors found that practitioners could benefit from using a flatter, “plateau-dome” rule with fixed interval individual blocks in the middle of the appointment session. Variations

of these rules have been found to perform well in a wide range of environments, including the presence of no-shows, non-punctual patients, non-punctual physicians, and different types of customers (e.g., new vs. returning). In this study, a set of fixed and variable interval rules are studied.

This study develops appointment scheduling rules for a multi-stage appointment scheduling environment. Multi-stage systems are common in both manufacturing and service environments. In such cases, the final output is a function of how the various stages interact and perform. The key difference is that the performance measures are now evaluated at more than one stage of service. Patient waiting time and physician idle time can now occur before each stage of service. In a single stage system, the waiting time of the patient is usually defined in terms of pre-service wait either from appointment time [9, 32] or from arrival time when patient unpunctuality is modeled [3, 8, 22]. The idle time of the physician is defined as the time the physician has between appointments (e.g., if he finishes with one patient early and the next patient has not yet arrived). Overtime involves any time that extends beyond the session end time. Service and arrival times in this environment are usually stochastic at each stage which further complicates the design of an efficient system [33].

In a multi-stage system, patients arriving for the first stage would be seen based on their appointment time. Patients for each subsequent stage are usually processed on a first-come-first-served basis into the second stage queue. However, there has been little attention paid to how schedules can be developed when the service is broken up into several stages [23]. A few prior studies have considered the effectiveness of MLSPs on patient wait times in various situations [25]. In emergency departments, the addition of an MLSP was found to result in a 12% increase in patient volume per shift with no overall reduction in waiting time [37]. In a similar study on emergency departments, [17] showed that wait times were less than half when a MLSP was present (12 min vs. 31 min). The integration of MLSPs in emergency departments can also improve patient flow [10]. In their study of a number of emergency departments, they found that patients were 1.6 times more likely to be seen within the wait time benchmarks than when MLSPs were not present. [27] considered psychiatric waiting times in days. They found a drop in waiting time from 32.5 days to 22.5 days after adding an MLSP.

There have been relatively few studies that address the pre-scheduled multi-stage outpatient environment. In [14], an outpatient clinic where patients proceed through multiple stages of service is considered. The authors use a simulation model to determine the best scheduling policy when there are multiple patient types. Their focus was on reducing the waiting time of patients from their appointment time to the start time of the first stage of service. In [23], the transient distribution of the multi-stage queue is used to show how optimal schedules that balance waiting

time of patients with idle time of the health care provider can be developed.

Appointment scheduling of outpatient surgical procedures in a multi-stage operating room is considered in [35]. The system had three stages, including pre-operation, surgery, and recovery. The authors propose several metaheuristic-based simulation approaches to solve the problem. Their modeling framework involved different types of patients, each of which has to be matched with a specific surgeon type. The objective was to minimize the waiting time of patients as well as the time the patient leaves the facility, and cancellations due to lack of time or resources. In [43], the relationship between appointment scheduling, capacity, and patient flow in multi-stage systems is examined. One of their findings was that there is a “sweet spot” when it comes to how many exam rooms to use. In the system they studied, fewer than three exam rooms greatly impeded patient flow and reduced performance, while more than three did not improve performance. They also found that scheduling low variance appointments first performed best.

Other work has focused on developing modeling frameworks and solution algorithms for improving the quality of service delivery in multi-stage systems. The question of how to allocate resources in such a way that the quality of service delivered is maximized subject to budgetary constraints is considered in [37]. In particular, the authors consider the problem from the perspective of customer perceptions of each stage of service. Their modeling framework is applied to a two-stage high-volume outpatient clinic where stage 1 involves patients checking in at a financial screening area (to determine health coverage eligibility) and stage 2 involves patients visiting the exam area where vital signs are checked and the patient is examined by their physician. The model is applied to several scenarios in order to determine how a budget could be allocated in order to improve patient perceptions of the quality of service.

Control charts have been used to measure the improvement at each stage [36, 44]. In [38], the authors base their study on the premise that performance at one stage of a process is statistically correlated with performance at the preceding stage. This study uses “cause selecting” control charts that account for the dependency between the various stages to monitor and identify potential problems in a multi-stage service process/system. The authors note that this approach is useful in indicating that problems may exist at specific stages of service. However, they cannot identify root causes of the problem and leave this issue for a clinic manager to determine.

### 3 Problem formulation

In this paper, appointment scheduling in a two-stage outpatient clinic with a single physician and a single MLSP is

considered. Performance is evaluated at each stage of service and service times of patients at each stage of service are stochastic. Arriving patients will either proceed to the MLSP for the first stage of service and the physician for the second stage or proceed to the physician directly. Patient service time distributions are assumed to be homogeneous (one distribution for the MLSP, and another for the physician). A simulation optimization model is developed to determine the best schedule under a variety of operating conditions.

### 3.1 Simulation optimization model

The problem examined is one of determining the start time of each patient’s appointment where the service times for each patient at each stage of service are random variables. The notation for the model is as follows.

- $t_i^{mlsp}$  Appointment start time of patient  $i$  (MLSP+Physician patient) for  $i = 1, 2, \dots, N$ .
- $t_j^{doc}$  Appointment start time of patient  $j$  (Doctor Only patient) for  $j = 1, 2, \dots, M$
- $S_i^{mlsp}$  Service time of patient  $i$  with MLSP
- $W_i^{mlsp}$  Waiting time of patient  $i$  for MLSP
- $I_k$  Idle time of doctor between patient  $k$  and  $k-1$  where  $k = 1, 2, \dots, N + M$
- $T$  Planned end time of clinic session
- $O$  Overtime of the clinic
- $c_w^{mlsp}$  Cost coefficient for patient’s waiting for the MLSP
- $c_w^{doc}$  Cost coefficient for patient’s waiting for the doctor
- $c_I$  Cost coefficient for doctor idle time
- $c_O$  Cost coefficient for clinic overtime
- $A_k$  Time the  $k^{th}$  patient arrives at the second stage (arrival time of the  $k$ th patient at the doctor) =  $\{t_1^{mlsp} + S_1^{mlsp} + W_1^{mlsp}, t_2^{mlsp} + S_2^{mlsp} + W_2^{mlsp}, \dots, t_N^{mlsp} + S_N^{mlsp} + W_N^{mlsp}, t_1^{doc}, t_2^{doc}, \dots, t_M^{doc}\}$  and  $A_{(k)}$  is the  $k$ th order statistic for  $k = 1, 2, \dots, N + M$
- $W_k$  Waiting time of the  $k^{th}$  patient seen by the doctor where  $W_{(k)}$  is the  $k$ th order statistic for  $k = 1, 2, \dots, N + M$
- $S_k$  Service time of the  $k^{th}$  patient seen by the doctor where  $S_{(k)}$  is the  $k$ th order statistic for  $k = 1, 2, \dots, N + M$

The following definitions apply.

$$W_1^{mlsp} = 0 \tag{1}$$

$$W_i^{mlsp} = \max \{W_{i-1}^{mlsp} + S_{i-1}^{mlsp} - t_i, 0\} \text{ for } i = 2, \dots, N \tag{2}$$

$$W_1 = 0 \text{ for } k = 1 \tag{3}$$

$$W_k = \max \{W_{k-1} + S_{k-1} - A_{k-1}, 0\} \text{ for } k = 2, \dots, N + M \tag{4}$$

$$I_1 = 0 \tag{5}$$

$$I_k = \max \{A_k - (A_{k-1} + W_{k-1} + S_{k-1}), 0\} \text{ for } k = 2, \dots, N + M \tag{6}$$

$$O = \max \{A_k - (A_{k-1} + W_{k-1} + S_{k-1}), 0\} - T \text{ for } k = N + M \tag{7}$$

$$\text{Flow Time of Patient } i = W_i^{mlsp} + W_k + S_i^{mlsp} + S_k \tag{8}$$

The objective is to determine a schedule that will minimize the following function, which represents: the expected total cost of patient waiting time at each stage of service, physician idle time, and clinic overtime.

$$\min_{t_i, t_j} c_w^{mlsp} E \left[ \sum_{i=1}^n W_i^{mlsp} \right] + c_w^{doc} E \left[ \sum_{j=1}^m W_k \right] + c_I E \left[ \sum_{k=1}^K I_k \right] + c_O [O] \tag{9}$$

subject to

$$t_1 \geq 0 \tag{10}$$

$$t_N \leq T \tag{11}$$

$$t_1^{mlsp} \leq t_2^{mlsp} \leq \dots \leq t_N^{mlsp} \tag{12}$$

$$t_1^{doc} \leq t_2^{doc} \leq \dots \leq t_M^{doc} \tag{13}$$

$$t_i, t_j \text{ integer} \tag{14}$$

Based on the above, it is always true that the total work for the physician is equal to  $T - \sum_j I_j + O$ . It is assumed that ap-

pointment start times are integer values. Patients may arrive early, late, or on time for their scheduled appointment. Each patient is seen by the MLSP in order of their appointment time unless a patient has not yet arrived and a patient with a later appointment is present. The order of the physician queue is based on both their *initial* appointment time and exam room availability. Patients in this queue are seen on a first-come first-serve basis. For instance, an MLSP patient will enter the physician queue only after the MLSP is finished. If a later-scheduled Physician-Only patient arrives before the MLSP is finished, they may be seen by the physician first. However, a Physician-Only patient will enter the physician queue only if there is an empty exam room.

Upon arrival all patients enter the exam room queue. A patient is either assigned to an exam room if one is available or waits in the queue. They simultaneously join either the MLSP or physician queue. Thus, especially in scenarios where the number of exam rooms is limited, MSLP patients that are in an exam room will often be seen by the physician before Physician-Only patients who have yet to be assigned a room. While total service times are generally higher with multiple stages, the goal is to determine schedules that reduce total

waiting time in the system for patients, while minimizing the idle time and overtime of the health care facility. This may have a positive effect on flow time (the total time the patient spends in the clinic). If waiting times decrease by more than the service times increase, this will result in reduced flow time for the patient.

While Eq. (9) includes the idle time of the physician, MLSP idle time is not included to allow for a direct comparison of performance for different proportions of patients allocated to the MLSP. It is also assumed that there are other “flexible” tasks the MLSP can do when not serving patients which will not have a noticeable or significant impact on the MLSP’s availability to serve patients (i.e., the MLSP is always available). They are able to pre-plan when and what other tasks they engage in based on the schedule and expected arrival times of patients. Similarly to prior single-stage studies, we assume there is a receptionist that brings patients to empty exam rooms. An additional benefit is patients spend more total time with medical care providers. In the outpatient care environment, patients often value having more time with service providers because the time with medical professionals is relatively short.

Given that the objective function and constraints in this model are stochastic functions, a simulation optimization approach is used to determine solutions for the problem. Simulation optimization is a stochastic optimization technique which is suitable for problems such as the one in this study since it is able to search for good solutions while simultaneously accounting for the complexity in the problem environment and the multiple sources of uncertainty present. This approach has been shown to produce good solutions for appointment scheduling problems in outpatient clinics [21, 22].

The simulation optimization problem can generally be defined by determining the vector of appointment start times that minimizes (9) [1, 11]. In this study, the optimization model is built in OptQuest [29]. The embedded algorithm combines scatter search and tabu search heuristics to search for feasible solutions and includes a neural network component to improve search efficiency. The population based heuristic iteratively generates sets of values for the decision variables in the problem which are then evaluated using simulation. The best solution for the problem in this case is the mean of the performance measure (9) from each iteration. The simulation optimization approach used in this study can be summarized as follows [1, 11, 29]:

#### *Step 1: Initialization*

An initial population of candidate solutions is generated to solve the following general problem  $\min f(\theta) = E[\gamma(\theta, \omega)]$ .  $t$ .  $\theta \in \Theta$  where  $\gamma$  is the sample performance measure,  $\theta$  is the vector of input variables with an upper and lower bound for

each input factor,  $u_i$  and  $l_i$ , and  $\omega$  represents a simulation replication.

#### *Step 2: Simulation*

A sufficient number of replications is performed for each candidate solution,  $\theta$ , to guarantee a stable solution. A neural network accelerator determines the number of replications required for computational efficiency. Based on results from initial experimentation, the number of replications was set at 500 for each candidate solution to ensure that a large sample of clinic environments were tested.

#### *Step 3: Optimization*

Feasible solutions are combined to create new solutions by generating linear combinations of prior solutions. Diversity is maintained by using both high and low quality solutions. Infeasible solutions are mapped to a feasible solution using a mixed-integer linear programming technique that minimizes the absolute deviation between the two points. Tabu memory functions prevent the algorithm from revisiting prior inferior solutions. The new set of candidate solutions is then evaluated using simulation (Step 2).

#### *Step 4: Stopping Criteria*

The algorithm can be stopped based on criteria specified by the user, such as projected or actual convergence (used in this study), after a number of heuristic iterations, or elapsed time.

### **3.2 Data collection**

Data was collected at a two stage outpatient clinic for 18 sessions which included both mornings and afternoons. A total of 285 usable observations were collected during these sessions. Approximately 70% of all patients were processed by a MLSP first before proceeding to see the physician. Patients who were not required to see the MLSP went directly to the physician. A summary of data is provided in Table 1.

For the first stage, the MLSP was found to have an average service time of 5.26 min with a standard deviation of 5.72; service times ranged from 0.37–36.6 min. These times were best fit to a lognormal distribution ( $\alpha = 0.05$ ). At the second stage, service time for the physician was divided into two categories of patients: those that saw both the MLSP and physician (MLSP+Physician) and those that saw only the physician (Physician-Only). For MLSP+Physician patients, physician service time was lower on average and less variable than for the Physician-Only patient. Waiting times for patients at each stage of service were also calculated. Average waiting time for the MLSP was 8.65 min with a standard deviation of

**Table 1** Statistics from observed clinic

	Average	Standard deviation	Min	Max
MLSP+Physician patients				
Service Time: MLSP	5.26	5.72	0.37	36.60
Service Time: Physician	9.26	5.77	0.93	46.73
Waiting Time for MLSP	8.65	12.34	0.00	71.07
Waiting Time for Physician	6.95	6.19	0.02	40.57
Physician only				
Service Time: Physician	11.66	7.41	0.63	38.77
Waiting Time for Physician	12.67	8.74	0.78	43.85

12.34 min. Waiting time for the physician was lower on average and less variable for MLSP+Physician patients.

Overall, *total* service time, waiting time, and flow time were on average 2.86, 2.93 and 5.79 min higher, respectively, for patients that went through both stages of service in this clinic. However, the benefit of having the MLSP is that the physician spent an average of 2.4 min less with each patient. If this were continued through an eight hour day, this would equate to an increase in physician capacity of 10.7 more patients on average.

### 3.3 Experimental design

The simulation optimization experiments are designed to determine how a number of factors impact performance and scheduling patterns. The values for the input parameters of the simulation optimization model are based on the data collected for this study and data and findings from earlier studies. In order to develop more general results, the input parameter values are varied in order to generate different clinic environments. Five factors are considered to determine their impact on system performance: cost structure, percentage of patients who see the MLSP, reduction in physician service time, clinic size, and method of measuring client waiting. In addition, we test some scenarios with and without a constraint on the latest an appointment can be scheduled.

Since many clinics consider the value of physician idle time and clinic overtime to be higher than patient waiting time [4, 21], the cost coefficient for overtime ( $c_o$ ) is tested at 1, 10, 30, and 50 in order to determine the impact on best schedule as costs incurred by the clinic change. The cost coefficients for waiting ( $c_w^{mlsp}$  and  $c_w^{doc}$ ) are set at 1. In some cases, physicians may be more concerned about overtime than idle time, since idle minutes can often be filled with other tasks. To represent this, the cost coefficient for idle time ( $c_l$ ) is set at zero [22, 39] for this first set of experiments. To determine the extent to which a MLSP improves performance, the percentage of patients who see the MLSP is tested at five levels: 0%, 25%, 50%, 75%, and 100%. It is assumed that all patients see the

physician and there is no constraint on the number of exam rooms available for this first set of experiments.

Reductions in the average time required for physician service are also analyzed. The clinic observed had a reduction in physician time of approximately 10% when an MLSP saw the patient first. In order to analyze a range of possible clinic environments where the MLSP provides more of the service, reduced service durations for the physician of 10%, 30% and 50% are tested, maintaining a coefficient of variation of 0.6 which has been used in prior studies [5]. Two clinic sizes of 12 and 24 appointments are modeled to determine if number of patients seen per session has an impact on the best schedule. These clinic sizes are based on the planned clinic duration and the expected total service times for patients. Service times for the physician and MLSP are assumed to follow a lognormal distribution; this is based on the data collected for this study as well as prior research (e.g., [22]). For the 24 appointment session, mean service durations tested for the physician are  $\text{LogN}(10,6)$  when the MLSP is not needed, and  $\text{LogN}(9, 5.4)$ ,  $\text{LogN}(7,4.2)$ , and  $\text{LogN}(5,3)$  minutes in order to determine the impact of reduced physician service times resulting from the presence of an MLSP in the system. MLSP service times are modeled as  $\text{LogN}(5,5)$ . For the 12 appointment system, in order to match session length and compare the systems fairly, all means and standard deviations for both servers are doubled.

Patient unpunctuality is modeled where the arrival of each patient is specified as their appointment time plus unpunctuality:  $A_i = t_i + \text{Normal}(-10, 15)$ . Earlier studies that included patient unpunctuality in a single stage environment propose that the best schedule depends on whether patient waiting is measured from the start time of their appointment or from the time of their arrival if they arrive early. For example, [22] show that if waiting is measured from the appointment start time, the plateau-dome scheduling rule usually performs best. However, if waiting time is measured from the time of a patient's arrival, an increasing interval & clustering rule (ICR), where intervals and block size increase towards the end of the session) is best for mitigating the effects of unpunctuality. Both measures of waiting time are tested in this study.

All levels of all factors are tested with each other. This results in a total of 320 clinic scenarios. The planned session time is four hours, which is representative of many outpatient clinics [4]. Actual sessions may be longer due to overtime.

## 4 Results and analysis

### 4.1 Simulation optimization results

In this section, we present the numerical results of the experiments. An important result is that average patient flow time is shorter when a MLSP is present even though total service

times increase when a patient sees both the MLSP and physician. This applies to all levels of MLSP utilization, including 100% of patients going to the MLSP. The difference is small (1.22% on average) but confirms that a clinic can add a MLSP without adversely affecting the time a patient spends in the clinic. Reducing patient flow time also has the positive effect of reducing overtime.

### 4.1.1 Clinic size

The impact of clinic size is similar to prior results for single server systems. Overall the smaller, 12 appointment clinics perform better. With fewer patients in the system, there is less variability and congestion even though the session length remains the same (also shown in [19]). This results in reduced total waiting time for patients and physician idle time by approximately 6.8%. However, overtime is consistently higher for smaller clinics. Smaller clinics tend to have longer appointment intervals and any delay that occurs near the end of a session results in more overtime. Thus, if an appointment is scheduled late and service time is longer, there is more chance of incurring overtime. This differs from single stage systems, where overtime and idle time of the physician increase and decrease together.

In general, larger clinics are more interesting to study because they are more complex. They have more patients moving through the system, provide more options for scheduling patterns, and require more careful management. Thus, subsequent discussion will focus primarily on the larger clinics.

### 4.1.2 Patient unpunctuality and waiting time measures

Earlier studies that included patient unpunctuality in a single stage environment found that the best schedule depends on whether patient waiting time is measured from appointment time or arrival time if they arrive early [22]. Those results are at least partially supported when a MLSP is added to the system. The difference is that a MLSP mitigates the differences between the two waiting measures. Comparing the best schedules developed under both measures, there is only a negligible difference in physician idle time and clinic overtime; idle time and clinic overtime are on average slightly worse (7.1% and 8.0% higher, respectively) when schedules are designed based on measuring from patient arrival time. This demonstrates that a clinic can build schedules based on either measure without adversely affecting the provider measures; it can decide which measure to use based on which one more accurately reflects their patients and their clinic [22]. Representative results for idle time and overtime are shown in Table 2.

The results in Table 2 are similar across the patient waiting time measures. However, they differ in terms of the pattern of appointments. When waiting is measured from arrival time,

**Table 2** Comparison of patient waiting time measures

% of patients that see MLSP		Waiting time measured from			
		$t_i$ (app't)		$A_i$ (arrival)	
		$c_o = 10$	$c_o = 30$	$c_o = 10$	$c_o = 30$
0%	Physician Idle Time	18.78	9.70	15.12	13.90
	Clinic Overtime	14.86	8.83	12.62	11.68
50%	Physician Idle Time	23.03	14.20	29.45	17.43
	Clinic Overtime	8.70	6.29	13.08	6.49
100%	Physician Idle Time	39.19	23.65	38.73	28.34
	Clinic Overtime	13.24	4.76	13.88	6.88

the increasing interval and clustering rule (IICR) is strongly supported in the 24 appointment case, and weakly supported in the 12 appointment case. However, the addition of the MLSP “levels” the schedule somewhat; the intervals in the IICR are not as extreme as in the single-stage case. When waiting is measured from appointment time, the plateau-dome rule is best, especially in the 24 appointment case. Thus, earlier results from prior studies are supported more strongly for larger clinics.

### 4.1.3 Cost structure

Similar to prior studies, as the cost coefficient for overtime increases, the best rule results in less overtime; patients are scheduled earlier, which in turn results in longer patient waiting times and less physician idle time. However, not all levels tested provided useful schedules. The results suggest that  $c_O = 1$  and  $c_O = 50$  are both extreme cases. Setting  $c_O = 1$  does not result in rules that adequately mimic the real world. For instance, when the last appointment is constrained to be no later than the end of the session, up to 5 appointments are scheduled at the very end of the session, resulting in higher idle time and excessive overtime that averages more than 35 min. Spreading appointments out (i.e., longer slots) has been shown to reduce patient waiting [21]. Thus, allowing appointment start times past the planned end time of the clinic is considered to see whether overall performance can be improved. When constraint (11) is removed, a number of appointments are set to start after the scheduled closing time, which reduces customer waiting. However, overtime averages between 75 and 110 min. Results for  $c_O = 50$  are similar to those for  $c_O = 30$  in both performance and scheduling patterns developed. For the discussion below, the results from  $c_O = 10$  and 30 are highlighted, since they result in more realistic representations of the systems under study. Results comparing cost levels are provided in Tables 2, 3, and 4, and Fig. 2.

**Table 3** Comparison of performance

Patient waiting time measured from time of appointment		Patient wait for MLSP (actual change in $W_i^{mlsp}$ )		Patient wait for Physician (% increase in $W_k$ )		Total patient wait (% increase)		Physician idle time (% increase)		Clinic overtime (% increase)		Objective Function (% increase)	
% of patients that see MLSP	% of patients that see MLSP	Patient wait for MLSP (actual change in $W_i^{mlsp}$ )	Patient wait for MLSP (actual change in $W_i^{mlsp}$ )	Patient wait for Physician (% increase in $W_k$ )	Patient wait for Physician (% increase in $W_k$ )	Total patient wait (% increase)	Total patient wait (% increase)	Physician idle time (% increase)	Physician idle time (% increase)	Clinic overtime (% increase)	Clinic overtime (% increase)	Objective Function (% increase)	Objective Function (% increase)
Patient waiting time measured from time of arrival													
0		0.00	0.00	8.3%	8.3%	2.4%	2.4%	Lowest	Lowest	70.8%	70.8%	8.5%	8.5%
25		15.72	0.00	Lowest	18.5%	Lowest	Lowest	24.4%	Lowest	58.1%	58.1%	4.1%	4.1%
50		32.41	15.52	6.2%	Lowest	11.6%	Lowest	22.7%	46.2%	Lowest	Lowest	Lowest	Lowest
75		52.75	34.85	3.4%	9.8%	16.0%	14.6%	56.1%	46.4%	4.9%	32.1%	4.1%	-3.7%
100		81.09	56.42	2.9%	6.6%	25.4%	17.3%	108.7%	97.1%	52.1%	9.7%	20.8%	Lowest
0		0.00	88.97	18.5%	22.1%	13.5%	40.9%	Lowest	143.9%	85.5%	Lowest	-15.9%	-12.4%
25		15.52		Lowest		Lowest		46.2%		88.0%		-8.0%	
50		34.85		9.8%		14.6%		46.4%		32.1%		-3.7%	
75		56.42		6.6%		17.3%		97.1%		9.7%		Lowest	
100		88.97		22.1%		40.9%		143.9%		Lowest		-12.4%	
Patient waiting time measured from time of arrival													
0		0.00	0.00	64.5%	64.5%	35.6%	35.6%	Lowest	Lowest	29.4%	29.4%	26.6%	26.6%
25		7.62	7.62	41.1%	41.1%	18.5%	18.5%	51.4%	51.4%	35.7%	35.7%	15.3%	15.3%
50		23.84	23.84	13.0%	13.0%	Lowest	Lowest	94.8%	94.8%	34.0%	34.0%	1.3%	1.3%
75		54.36	54.36	11.7%	11.7%	7.7%	7.7%	103.1%	103.1%	Lowest	Lowest	Lowest	Lowest
100		91.87	91.87	Lowest	Lowest	8.9%	8.9%	156.2%	156.2%	42.2%	42.2%	9.6%	9.6%
0		0.00	0.00	28.8%	28.8%	4.0%	4.0%	Lowest	Lowest	142.2%	142.2%	31.2%	31.2%
25		8.49	8.49	22.7%	22.7%	0.9%	0.9%	16.2%	16.2%	95.5%	95.5%	18.3%	18.3%
50		30.23	30.23	15.7%	15.7%	Lowest	Lowest	25.4%	25.4%	34.6%	34.6%	3.6%	3.6%
75		62.16	62.16	14.4%	14.4%	6.0%	6.0%	32.0%	32.0%	Lowest	Lowest	Lowest	Lowest
100		108.73	108.73	Lowest	Lowest	4.5%	4.5%	103.9%	103.9%	42.7%	42.7%	8.8%	8.8%

\*24 appointments, physician service time for patients that see the MLSP =  $\text{LogN}(9,5.4)$



**Table 4** Impact of reducing the physician's service times (percentage change)\*

	Physician service times	Total patient wait*	Physician idle time	Clinic overtime	Objective function
$c_{co} = 10$	LogN(9,5.4)				
	LogN(7,4.2)	-13.3%	13.4%	-39.5%	-21.8%
	LogN(5,3)	-23.1%	36.8%	-69.9%	-38.2%
$c_{co} = 30$	LogN(9,5.4)				
	LogN(7,4.2)	-22.3%	50.4%	-37.7%	-28.7%
	LogN(5,3)	-33.0%	88.1%	-67.2%	-47.3%

\*% that go to MLSP = 25, patient wait is from time of appointment

#### 4.1.4 Percentage of patients allocated to the MLSP

In cases where a patient does not see the MLSP (i.e., a Physician-Only patient), their service time at the physician is longer than if an MLSP is seen first. Table 3 provides a comparison of performance when  $c_o = 10$  and  $c_o = 30$ . When  $c_o = 10$ , performance is best when approximately 50% of patients are scheduled to see the MLSP first (75% if waiting is measured from arrival time). As the proportion of patients that see the MLSP increases, there are some overall patterns evident. First, patient waiting time for the MLSP increases. This is consistent with the well-known nonlinear relationship between resource utilization and waiting time in the queuing literature. An increase in the number of patients utilizing the services of the MLSP results in more congestion and longer queues at this stage.

Second, patient waiting time for the physician varies based on how it is measured. If it is measured from arrival time of the patient, waiting time falls. Appointments are scheduled later which reduces the waiting between MLSP and physician as the proportion allocated to the MLSP increases. If measured from appointment time, there is no consistent pattern. The lowest patient waiting time occurs when 25% of patients see the MLSP. As the proportion allocated to the MLSP increases from 25% to 100%, performance is not affected by patients showing up early for their appointments. Every appointment is scheduled earlier on average (more than 2 min earlier for the scenarios in Table 3). If the MLSP is free before the scheduled appointment time and the patient is present, the patient will be seen at that time, after which they wait for the physician. As more patients see the MLSP, the system becomes congested with more patients waiting to see the physician. In terms of total patient waiting time at both stages of service, this is minimized when 25–50% of patients see the MLSP. When compared to a system without an MLSP, more work per patient is required in total, but total waiting time is also reduced. Thus, reducing the service time for the primary caregiver, even by a small amount, can improve system flow. However, as utilization of the MLSP increases, congestion for that stage results in higher overall waiting time.

Third, physician idle time increases as more patients see the MLSP because there is less work and less variability in that work. Thus, adding a MLSP can be valuable in providing the physician more time to see other patients or performing other tasks. In most cases, when 100% of patients see the MLSP, the congestion experienced by the MLSP has a large negative impact on physician idle time.

Finally, in most cases, overtime is lowest when 50–75% of patients see the MLSP. For example, for  $c_o = 30$  and waiting is measured from the time of appointment, overtime is lowest at 100% (because patients are scheduled early to minimize overtime). Overall performance is best when 75% of patients are directed to an MLSP first. In all cases there is very little difference in performance for an allocation strategy of 50% and 75%. This is supported by the data for overtime. In most cases, overtime is minimized at a proportion of 50% or 75%, suggesting that these levels result in better patient flow through the stages than either higher or lower proportions.

#### 4.1.5 Reduction of physician service times

Another factor considered is the benefit from further reductions in physician service time for MLSP+Physician patients resulting from shifting work to a highly trained MLSP. Service times for Physician-Only patients remain the same. Representative results are provided in Table 4, where results measuring waiting time from appointment time and a 25% allocation of patients to the MLSP are given. Results for all other levels and for measuring waiting from arrival time are similar in all respects. For the same proportion of patients that see the MLSP, total waiting time and waiting time for the physician are reduced considerably with no change in waiting time for the MLSP. Physician idle time increases, clinic overtime decreases considerably (because the bottleneck resource has less work), and overall performance is improved. This suggests that if adding an MLSP can reduce physician service times, more patients could be scheduled without adversely affecting overall performance.

Tables 3 and 4 highlight a key difference between single stage and multi-stage clinics. In a single stage setting, there is a consistent tradeoff between patient waiting time and both

physician idle time and clinic overtime. A scheduling rule or policy that reduces waiting time (and thus patient flowtime) will result in increased idle time and overtime, and vice versa. This tradeoff is so consistent that it has often been analyzed and interpreted by the use of an efficient frontier (e.g., [5, 15]). When a MLSP is added there are two waiting periods ( $W_i^{mlsp}$  and  $W_k$ ) with multiple inter-related queues and results show that this tradeoff is no longer consistent. For example, a schedule that reduces waiting time for patients will sometimes result in lower overtime for the clinic. In addition, overtime and idle time do not always increase or decrease together. Results are more dependent on the percentage of patients that see the MLSP – although there are no consistent patterns except that physician idle time consistently increases as the percentage that see the MLSP increases.

In terms of basic queuing principles, an increase in resource utilization results in an increase in average flow time. In this problem, the point at which average flow time for patients starts to increase is when the allocation of patients to the MLSP increases beyond 75%. At allocation levels between 25 and 75%, however, the corresponding decrease in physician service time sometimes results in a reduction in flow time for patients and a corresponding reduction in overtime for the clinic.

### 4.2 Scheduling policies

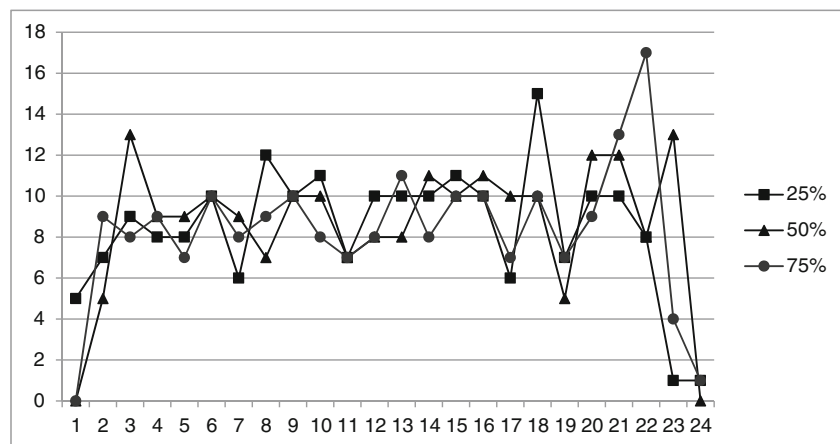
The best schedules found using simulation optimization may be difficult to implement from a practical standpoint. Therefore, several rules were devised based on the simulation optimization results that may function better in terms of practical implementation and that are more generalizable to a variety of clinic environments. Figure 1 provides an example of some of the best schedules found using simulation optimization. As

shown, these schedules followed an approximate, but not perfect, plateau-dome pattern. From the standpoint of practical implementation, “smoothing” out the time slots in the middle of the schedule such that the intervals are of equal length would be preferable. As such, dome and plateau-dome rules were developed based on the appointment interval lengths and block sizes from the simulation optimization results.

In addition, in order to ensure that other rule variations are also compared, several rules that have been shown to be best in prior literature (on single stage studies) and are common in practice were tested. These are the single-block fixed interval (SBFI) and multi-block fixed interval (MBFI) patterns. These patterns were less common in the simulation optimization results. However, general rules with these properties were devised based on the intervals and block sizes found in the simulation optimization study and tested to determine if they perform well in this multistage environment. Detailed numerical examples of each rule are in Appendix Table 8. These rules are tested in simulation experiments, described below.

Since there are two different types of patients (MLSP+Physician and Physician-Only), sequencing rules are also proposed. Based on the simulation optimization results, it is advisable to spread the two types of patients throughout the session, keeping servers at both stages consistently busy. Four patterns that alternate patient types are tested. In order to provide initial work for both servers, both a Physician-Only and MLSP+Physician, are scheduled at the beginning of the session. Appendix Table 9 summarizes the sequencing patterns. The number of exam rooms required to ensure good performance and different cost structures are also tested in these simulation experiments. The previous model imposed no constraints on the number of exam rooms available. It was assumed that patients would be seen in order of their arrival into the MLSP and Physician queues. However, if there are restrictions on the number of exam rooms, an arriving

**Fig. 1** Examples of schedules from the simulation optimization\*. \* $c_{co} = 30$ , patient wait is from time of arrival, for all 5 levels of % that go to MLSP



**Table 5** Factors and levels for simulation experiments

Factors	Levels
Cost Structure ( $c_I, c_O$ )	(1,1), (5,5), (15,15)
# of exam rooms	2,3,4,5
Rules	SBFI, MBFI, PD1, PD2, PD3, PD4, Dome
Sequencing	Alt1, Alt2, Alt3, Alt4

patient may have to wait even if their resource is free. For example, an arriving Physician-Only patient would have to wait even if the doctor is free if all the exam rooms are currently occupied (this occurs very rarely). All sequences were tested for all rules and exam room levels. For the remaining factors, the best levels from the earlier experiments were used. The results are analyzed with a full factorial ANOVA model and follow-up tests.

**4.2.1 Simulation experiments**

The scheduling rules in Appendix Table 8 are simulated using the performance measure in (9). The clinic size is 24 patients, the physician service time for patients that see the MLSP is  $\text{LogN}(9,5.4)$ , and waiting is measured from appointment time since this definition is more common in the literature and most clinics consider waiting costs to start at this time. In [22], it was demonstrated that physician idle time and clinic overtime have a similar effect on performance and scheduling rule development. This is because both produce rules that favor the physician and thus, have similar effects. Their results also suggest that when both are included in the performance measure, there is a cumulative effect. For instance, if  $c_I$  and  $c_O$  are both set at 5 in (9), the resulting performance and schedules are similar to using only *either* idle time *or* clinic overtime in

the measure and setting the cost factor to ten. Thus, to evaluate scheduling rules,  $c_I$  and  $c_O$  are set at 5 in order to parallel the cases in earlier experiments where  $c_O = 10$ , and 15 to parallel the cases where  $c_O$  was set at 30.

Table 5 shows that 336 scenarios are used for the full factorial experiment. In addition, the percentage of patients seeing the MLSP was run and analyzed at all five levels. Results are reported for the case where 50% of patients are scheduled to see the MLSP, since this was shown to be best for all rules. Each scenario is run for 1000 replications, using common random numbers to reduce variation due to that input.

**4.2.2 ANOVA results**

An ANOVA using Eq. (9) as the dependent variable, with  $c_I$  and  $c_O$  at 5, shows that Rule and Number of Rooms were significant but the main effect of Sequence was not. In addition, the only interaction that is significant is Number of Rooms and Sequence, so it is significant only due to the Number of Rooms. The ANOVA results with  $c_I$  and  $c_O$  at 15 were similar, except none of the interactions are significant (Table 6). The lack of significance for sequence indicates clinics can choose to alternate every patient or every second patient.

A number of follow-up tests were done to understand the significance of Number of Rooms and Rules. Tukey HSD and Sheffe tests showed that having only two exam rooms resulted in significantly worse performance than having three, four, or five rooms. There is no statistical difference among the latter (see Table 7) which is in line with the findings from [43]. This indicates that at times when both servers are busy, performance is maximized if the next patient can already be prepped and waiting in a room. These results are in line with [43],

**Table 6** ANOVA for  $c_I=15$  and  $c_O = 15$

Source	Type III sum of squares	df	Mean square	F	Sig.
Corrected model	442032282.0	111	3982272.81	24.57	0.000
Intercept	63877579673.0	1	63877579673.04	394073.41	0.000
Rule	223369490.7	6	37228248.45	229.67	.000
Sequence	660782.3	3	220260.78	1.36	.253
NumRooms	212831260.7	3	70943753.55	437.67	.000
Rule * Sequence	895519.8	18	49751.10	0.31	.998
Rule * NumRooms	2758581.9	18	153254.55	0.95	.522
Sequence * NumRooms	1154839.3	9	128315.48	0.79	.624
Rule * Sequence * NumRooms	361807.4	54	6700.14	0.04	1.000
Error	18136556393.6	111888	162095.63		
Total	82456168348.6	112000			
Corrected total	18578588675.6	111999			

**Table 7** Tukey HSD and Scheffe tests: Number of rooms

NumRooms	N	Subset	
		1	2
Tukey HSD <sup>a,b</sup>	4	28000	431.669
	5	28000	432.287
	3	28000	434.041
	2	28000	482.591
	Sig.		.659
Scheffe <sup>a,b</sup>	4	28,000	431.669
	5	28,000	432.287
	3	28,000	434.041
	2	28,000	482.591
	Sig.		.724

Means for groups in homogeneous subsets are displayed. Based on observed means. The error term is Mean Square(Error) = 59648.165

<sup>a</sup> Uses Harmonic Mean Sample Size = 28000.00

<sup>b</sup> Alpha = .05

which also showed that no more than three exam rooms are required when there are two servers.

Based on the above results we focus on results for rules with three rooms. These results are shown in Fig. 2 (averaged over all sequencing levels). The error bars represent the sampling error ( $\alpha = 0.05$ ) associated with the mean values for the performance measure given in (9). These 95% confidence intervals average 1.80% across the rules, with a maximum of 2.10%.

Figure 2 suggests that when  $c_I$  and  $c_O$  are equally weighted with the cost of waiting, the PD2, PD3 and MBFI rules are best. When  $c_I$  and  $c_O$  equal 5, all rules exhibit similar performance, as verified with a Tukey HSD test. When  $c_I$  and  $c_O$  equal 15, the best rules are PD1 and PD4 where all other rules are significantly inferior. Regardless of cost structure, as the physician’s time becomes more important, the plateau-dome rules with shorter appointment slots perform best. Thus, the

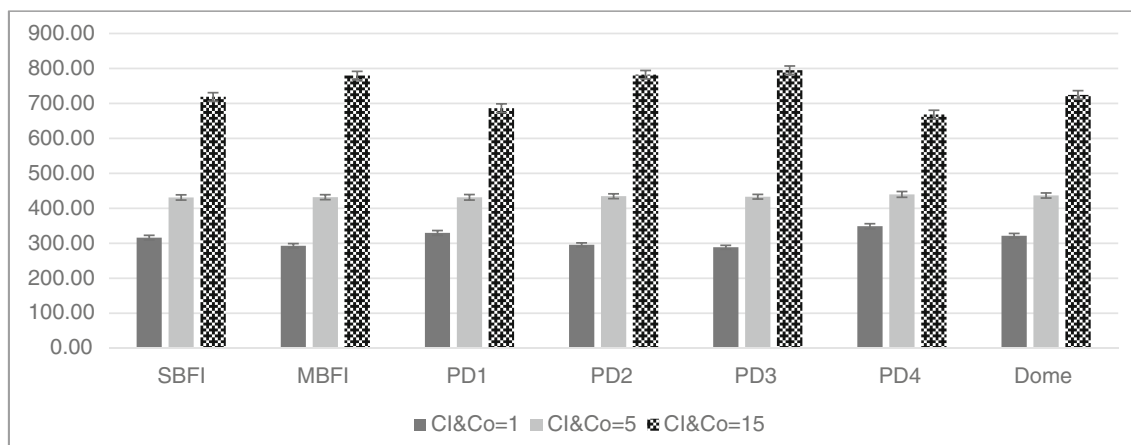
plateau-dome is a robust rule that applies in the multi-stage environment.

Figure 3 demonstrates the contribution of the components (without cost weightings) to the overall measure for this scenario. The figure reflects the 95% confidence intervals (average, maximum) for total waiting time for the MLSP (4.68%, 5.16%), total waiting time for the physician (1.90%, 1.99%), overtime (4.28%, 5.10%), and total physician idle time (2.76%, 3.23%). As explained above, idle time and overtime are best when the plateau is lower (i.e., appointment intervals are shorter). The 12 MLSP+Physician patients experienced a fairly short wait for the MLSP, while all 24 patients had a slightly longer wait on average for the physician. This is due to higher utilization of the physician. In addition, Physician-Only patients waited slightly longer on average for the physician than MLSP+Physician patients for most rules, since the latter are already in an exam room when the first stage of service is complete (see Figure 4). The figure reflects 95% confidence intervals of 2.76% (maximum of 2.91%) and 1.21% (maximum of 1.30%) on average for Physician-Only and MLSP patients, respectively.

### 5 Discussion

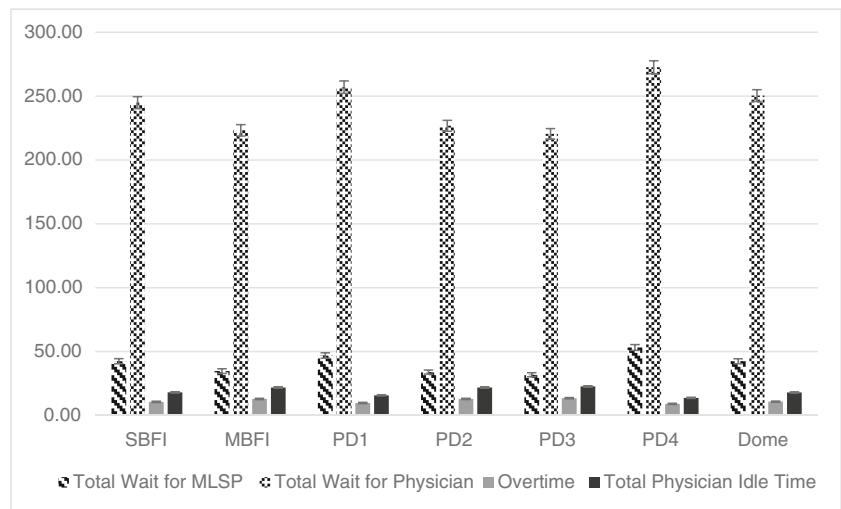
In this study, the impact of adding a MLSP to serve patients in a single stage outpatient clinic is studied. The purpose was to determine the circumstances under which a MLSP becomes the best strategy for a clinic and to determine how policies from a single-stage environment should be adjusted to accommodate the additional stage of service. Appointment scheduling rules are developed for a multi-stage service system using a simulation optimization approach. Based on the results, general scheduling policies are formulated and tested using simulation.

A key finding of this study is that the percentage of patients allocated to the MLSP has a large impact on performance. An



**Fig. 2** Comparison of rules based on the performance measure\*. \*Dependent variable: Eq. (9), 3 exam rooms, 50% of patients see the MLSP

**Fig. 3** Components in the performance measure\*. \*Dependent variable: Eq. (9), 3 exam rooms, 50% of patients see the MLSP



important decision in the context of a multi-stage clinic is the best strategy in terms of the proportion of patients to allocate to each stage of service. It was shown that overall performance is best if 50–75% of patients are allocated to the MLSP. Patients that see both providers require a longer total service time. However, the reduction of work for the physician is of greater benefit to the system, reducing overtime and flow time for patients. Performance is best when the MLSP sees some but not all patients.

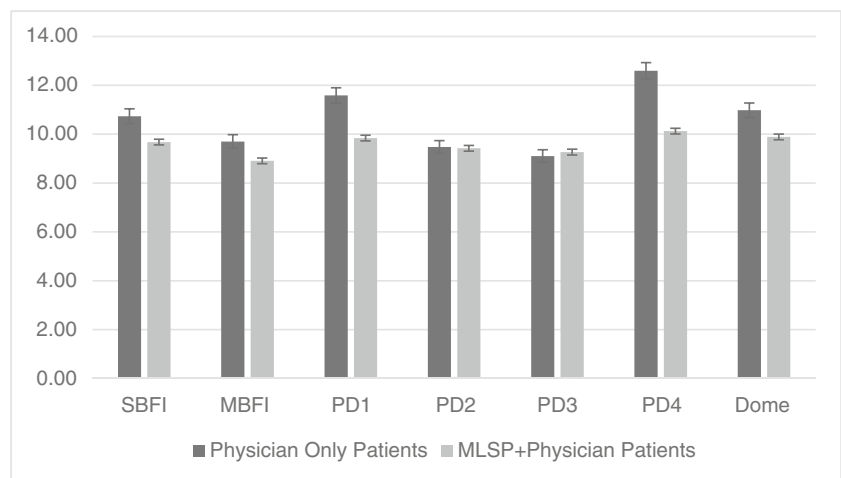
In many single-stage studies, an important tradeoff is highlighted, often presented in the form of an efficient frontier of scheduling rules or policies. A rule that reduces physician idle time or overtime consistently results in higher patient waiting time, and vice versa (e.g., [5, 15]). In a multi-stage system the tradeoff between waiting time and physician idle time or overtime is not consistent. Performance is dependent on the percentage of patients that see the MLSP. It was shown that as this percentage increases, physician idle time increases. However, waiting time and overtime do not follow a consistent pattern for different scenarios, often being lowest at the

25%, 50% or 75% levels. Thus, idle time and overtime of the physician do not increase and decrease in the same direction in a multi-stage system as they do in a single-stage system.

Results showed cost structure to be important in terms of the best scheduling policy. If the cost of overtime in the objective function is low (in this case, if it is equal to the waiting time cost coefficient), then realistic schedules cannot be produced. High values for the coefficient led to no appreciable difference in performance. For example, a coefficient of 50 (i.e., clinic overtime is 50 times more valuable than patient waiting) did not produce significantly different results than a coefficient of 30.

The effects of clinic size were also significant. It was shown that if waiting is measured from arrival time, the IICR (“increasing interval and clustering rule”) is strongly supported in the 24 appointment case and weakly supported in the 12 appointment case. The addition of the MLSP “levels” the schedule somewhat with appointment intervals that are not as extreme. When waiting is measured from appointment time, the plateau-dome rule performs best,

**Fig. 4** Average waiting time for the physician. \*Dependent variable: Eq. (9), 3 exam rooms, 50% of patients see the MLSP



especially in the 24 appointment case. Thus, earlier work in single-stage systems is supported more strongly for larger clinics.

If the MLSP is able to carry out even more tasks, further reducing the physician’s time with patients, then patient waiting, overtime and overall performance improve. The result is that the physician has more idle time which could be used to carry out paperwork and other tasks, or possibly to increase revenue by seeing more patients. Another possible benefit of seeing more patients beyond an increase in income is the ability to better serve the physician’s panel of patients. However, the “human” side of this requires more study. It may be difficult to “train” a physician to spend less time with patients (e.g., to shorten a typical 10 min appointment to 5 min). This may also lead to a perception of diminished service quality on the part of patients.

This study has demonstrated a number of other findings. It was shown that smaller clinics are easier to manage due to reduced variability. In addition, sequencing of Physician-Only and MLSP+Physician patients should alternate between these two types, but there is little difference if the alternating occurs with every patient or every second patient. It was also found

that three exam rooms are sufficient for a system with one physician and one MLSP.

Further work could be done to improve the understanding of multi-stage environments. As mentioned, the benefits and drawbacks of reducing the physician’s time with patients could be explored. Scenarios where some patients see only the MLSP and do not need to see the physician at all could be studied. Clinics with multiple physicians and heterogeneous patients with a single MLSP could also be the subject of future study. It is worthwhile noting that in such cases, coordination will be more difficult and temporary bottlenecks may be created when an MLSP patient arrives if the MLSP is busy completing work for another physician. Another potential area for research is to consider using the MLSP capabilities to help manage patient waiting in real time. When the clinic is busier, the MLSP does more of the work and when it is less busy the physician spends more time with patients. In addition, it would be useful to collect enough data to study the correlation between the MLSP and physician service times. This may address the question of whether a longer MLSP service time indeed results in a shorter physician service time and how it relates to the clinic structure.

## Appendix 1

**Table 8** Scheduling rules (Slot Lengths)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
SBFI: Single-block, Fixed Interval	0	0	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
MBFI: Multi-block, Fixed Interval	0	0	20	0	20	0	20	0	20	0	20	0	20	0	20	0	20	0	20	0	20	0	20	0
PD1: Plateau-Dome 1	0	0	9	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	9	7	5
PD2: Plateau-Dome 2	0	0	6	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	5	2	0
PD3: Plateau-Dome 3	0	0	8	10	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	8	6	5	3
PD4: Plateau-Dome 4	0	0	8	9	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	7	6	5
Dome	0	0	8	9	9	10	10	10	11	11	11	11	12	12	11	11	11	10	9	9	8	7	6	4

**Table 9** Sequencing Patterns\*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Alt1	P	M	P	M	P	M	P	M	P	M	P	M	P	M	P	M	P	M	P	M	P	M	P	M
Alt2	P	M	P	P	M	M	P	P	M	M	P	P	M	M	P	P	M	M	P	P	M	M	P	M
Alt3	P	M	M	M	P	P	M	M	P	P	M	M	P	P	M	M	P	P	M	M	P	P	M	P
Alt4	P	M	P	P	M	P	P	M	M	P	M	M	P	P	M	P	P	M	M	P	M	M	P	M

\*P = Physician-Only, M = MLSP+Physician

\*24 appointments, waiting measured from appointment time, 50% of patients see MLSP

## References

1. Andradóttir S (2006) Simulation optimization with countably infinite feasible regions: efficiency and convergence. *ACM Trans Model Comput Simul* 16(4):357–374. <https://doi.org/10.1145/1176249.1176252>
2. Bailey N (1952) A study of queues and appointment Systems in Hospital Outpatient Departments with special reference to waiting times. *J R Stat Soc* 14:185–199
3. Blanco White M, Pike M (1964) Appointment Systems in Outpatients' clinics and the effect on patients' unpunctuality. *Med Care* 2(3):133–145. <https://doi.org/10.1097/00005650-196407000-00002>
4. Cayirli T, Veral E (2003) Outpatient scheduling in health care: a review of literature. *Prod Oper Manag* 12:519–549
5. Cayirli T, Veral E, Rosen H (2006) Designing appointment scheduling Systems for Ambulatory Care Services. *Health Care Manag Sci* 1:47–58
6. Cayirli T, Veral E, Rosen H (2008) Assessment of patient classification in appointment system design. *Prod Oper Manag* 17:338–353
7. Cortada JW, Gordon D, Lenihan B (2012) The value of analytics in healthcare: from insights to outcomes. IBM global business services executive report. IBM Institute for Business Value, IBM Corporation, New York, USA
8. Cox T, Birchall J, Wong H (1985) Optimising the queuing system for an ear, nose and throat outpatient clinic. *J Appl Stat* 12(2):113–126. <https://doi.org/10.1080/02664768500000017>
9. Denton B, Gupta D (2003) A sequential bounding approach for optimal appointment scheduling. *IIE Trans* 35(11):1003–1016. <https://doi.org/10.1080/07408170304395>
10. Ducharme J, Alder RJ, Pelletier C, Murray D, Tepper J (2009) The impact on patient flow after the integration of nurse practitioners and physician assistants in 6 Ontario emergency departments. *J Can Assoc Emerg Physicians* 11:455–461
11. Fu MC (2002) Optimization for simulation: theory vs. *Prac INFORMS J Comput* 14(3):192–215. <https://doi.org/10.1287/ijoc.14.3.192.113>
12. Gonsalves T, Itoh K (2009) Service optimization with patient satisfaction in healthcare systems. *Journal of simulation, suppl. Spec Issue: Simul Healthc Part 1*:150–162
13. Gupta D, Denton B (2008) Appointment scheduling in health care: challenges and opportunities. *IIE Trans* 40(9):800–819. <https://doi.org/10.1080/07408170802165880>
14. Harper PR, Gamlin HM (2003) Reduced outpatient waiting times with improved appointment scheduling: a simulation modelling approach. *Oper Res-Spektrum* 25(2):207–222. <https://doi.org/10.1007/s00291-003-0122-x>
15. Ho C, Lau H (1992) Minimizing Total cost in scheduling outpatient appointments. *Manag Sci* 38(12):1750–1764. <https://doi.org/10.1287/mnsc.38.12.1750>
16. HRSA (2013) Projecting the supply and demand for primary care practitioners through 2020. U.S. Department of Health and Human Services, Health Resources and Services Administration, National Center for health workforce analysis, Rockville, Maryland. <http://bhpr.hrsa.gov/healthworkforce/supplydemand/usworkforce/primarycare/projectingprimarycare.pdf>
17. Janakiraman N, Meyer R, Hoch S (2011) The psychology of decisions to abandon waits for service. *J Mark Res* 48:970–984
18. Jennings N, O'Reilly G, Lee G, Cameron P, Free B, Bailey M (2008) Evaluating outcomes of the emergency nurse practitioner role in a major urban emergency department. *J Clin Nurs* 17(8):1044–1050. <https://doi.org/10.1111/j.1365-2702.2007.02038.x>
19. Klassen K, Rohleder T (1996) Scheduling outpatient appointments in a dynamic environment. *J Oper Manag* 14(2):83–101. [https://doi.org/10.1016/0272-6963\(95\)00044-5](https://doi.org/10.1016/0272-6963(95)00044-5)
20. Klassen K, Yoogalingam R (2008) An assessment of the interruption level of doctors in outpatient appointment scheduling. *Oper Manag Res* 1(2):95–102. <https://doi.org/10.1007/s12063-008-0013-z>
21. Klassen K, Yoogalingam R (2009) Improving performance in outpatient appointment services with a simulation optimization approach. *Prod Oper Manag* 18(4):447–458. <https://doi.org/10.1111/j.1937-5956.2009.01021.x>
22. Klassen K, Yoogalingam R (2014) Patient unpunctuality: strategies for appointment policy design. *Decis Sci* 45(5):881–911. <https://doi.org/10.1111/dec.12091>
23. Kuiper A, Mandjes M (2015) Appointment scheduling in tandem-type service systems. *Omega* 57:145–156. <https://doi.org/10.1016/j.omega.2015.04.009>
24. Laganga L, Lawrence S (2007) Clinic overbooking to improve patient access and increase provider productivity. *Decis Sci* 38(2):251–276. <https://doi.org/10.1111/j.1540-5915.2007.00158.x>
25. Lin C (2015) An adaptive scheduling heuristic with memory for the block appointment system of an outpatient specialty clinic. *Int J Prod Res* 53(24):7488–7516. <https://doi.org/10.1080/00207543.2015.1084060>
26. Maister D (1985) The psychology of waiting lines. <http://davidmaister.com/articles/the-psychology-of-waiting-lines/> Accessed 3 April 2013
27. Meats P, Ashton T (1997) Nurses' help in psychiatric outpatient clinics. *Psychiatr Bull* 21(11):677–679. <https://doi.org/10.1192/pb.21.11.677>
28. O'Hare S (2010) Mid-level providers in a changing healthcare workforce. *Becker's Hospital Review* August 17 <http://www.beckershospitalreview.com/compensation-issues/mid-level-providers-in-a-changing-healthcare-workforce.html> Accessed September 2016
29. OptTek Systems, Inc. <http://www.opttek.com> Accessed September 2016
30. Powers M (2011) Reducing patient wait times: Examine your operations to boost efficiency. *ENT Today* October 10. <http://www.enttoday.org/article/reducing-patient-wait-times-examine-your-operations-to-boost-efficiency>. Accessed September 2012
31. Rising E, Baron R, Averill B (1973) A system analysis of a university health service outpatient clinic. *Oper Res* 21:1020–1047
32. Robinson L, Chen R (2003) Scheduling doctors' appointments: optimal and empirically-based heuristic policies. *IIE Trans* 35(3):295–307. <https://doi.org/10.1080/07408170304367>
33. Rohleder TR, Lewkonja P, Bischak DP, Duffy P, Hendijani R (2011) Using simulation modeling to improve patient flow at an outpatient orthopedic clinic. *Health Care Manag Sci* 14:135–145
34. Salzarulo PA, Mahar S, Modin S (2016) Beyond patient classification: using individual patient characteristics in appointment scheduling. *Prod Oper Manag* 25:1056–1072
35. Saremi A, Jula P, ElMekawy T, Wang GG (2013) Appointment scheduling of outpatient surgical Services in a Multi-Stage Operating Room Department. *Int J Prod Econ* 141(2):646–658. <https://doi.org/10.1016/j.ijpe.2012.10.004>
36. Shi J, Zhou S (2009) Quality control and improvement for multi-stage systems: a survey. *IIE Trans* 41(9):744–753. <https://doi.org/10.1080/07408170902966344>
37. Soteriou AC, Hadjinicola GC (1999) Resource allocation to improve service quality perceptions in multi-stage service systems. *Prod Oper Manag* 8:221–239
38. Sulek JM, Maruchek A, Lind MR (2006) Measuring performance in multi-stage service operations: an application of cause selecting control charts. *J Oper Manag* 24:711–727
39. Vanden Bosch PM, Dietz DC, Simeoni JR (1999) Scheduling customer arrivals to a stochastic service system. *Nav Res Logist* 46(5):549–559. [https://doi.org/10.1002/\(SICI\)1520-6750\(199908\)46:5<549::AID-NAV6>3.0.CO;2-Y](https://doi.org/10.1002/(SICI)1520-6750(199908)46:5<549::AID-NAV6>3.0.CO;2-Y)

40. Vissers J (1979) Selecting a suitable appointment system in an outpatient setting. *Med Care* 17(12):1207–1220. <https://doi.org/10.1097/00005650-197912000-00004>
41. Wang P (1993) Static and dynamic scheduling of customer arrivals to a single-server system. *Nav Res Logist* 40(3):345–360. [https://doi.org/10.1002/1520-6750\(199304\)40:3<345::AID-NAV3220400305>3.0.CO;2-N](https://doi.org/10.1002/1520-6750(199304)40:3<345::AID-NAV3220400305>3.0.CO;2-N)
42. Wang P (1997) Optimally scheduling N customer arrival times for a single-server system. *Comput Oper Res* 24(8):703–716. [https://doi.org/10.1016/S0305-0548\(96\)00093-7](https://doi.org/10.1016/S0305-0548(96)00093-7)
43. White D, Froehle C, Klassen K (2011) The effect of integrated scheduling and capacity policies on clinical efficiency. *Prod Oper Manag* 20:442–455
44. Yang S-F (2011) Using cause selecting control charts to monitor dependent process stages with attributes data. *Expert Syst Appl* 38(1):667–672. <https://doi.org/10.1016/j.eswa.2010.07.018>