Proceedings of the 2016 Winter Simulation Conference
T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, eds.

A DECISION SUPPORT SYSTEM FOR REAL-TIME AND DYNAMIC SCHEDULING OF MULTIPLE PATIENT CLASSIFICATIONS IN AMBULATORY CARE SERVICES

William P. Millhiser

Loomba Department of Management
Zicklin School of Business, Baruch College
One Bernard Baruch Way
New York, NY 10010, USA

Emre A. Veral

Loomba Department of Management
Zicklin School of Business, Baruch College
One Bernard Baruch Way
New York, NY 10010, USA

ABSTRACT

We propose a methodology to provide real-time assistance for outpatient scheduling, involving multiple patient types. Schedulers are shown how each prospective placement would impact the day's operational performance for patients and providers. Rooted in prior literature and analytical findings, the information provided to schedulers about vacant slots is based on the probabilities that the calling patient, the already-existing appointments, and the session-end time will be unduly delayed. The information is dynamically updated after every new booking; calculations are driven by historical consultation times and no-show data, and a simulation tool that implements the underlying analytical methodology. Our findings lead to practical guidelines for constructing templates that provide allowances for different service time lengths and variability, no-show rates, and provider-driven performance targets for patient delays and providers' overtime. Extensions to OR scheduling are viable as avoiding session overtime and procedures' completion time delays involve similar considerations.

1 INTRODUCTION

The process of scheduling service appointments is a game in which customers negotiate with schedulers for time slots that satisfy each party's particular needs (Herrler *et al.* 2002). Schedulers engage in this negotiation with a partially-filled appointment book and "must schedule clients dynamically…as they receive calls without knowing what type of clients will call in the future for that session" (Klassen and Rohleder 1996, p. 84). The general scheduling literature calls such a process *online* scheduling, but the literature on ambulatory care management has given alternative scheduling names such as "dynamic" (Klassen and Rohleder 1996, Liu *et al.* 2010), "myopic and sequential" (Muthuraman and Lawley 2008), and simply "sequential" (Turkcan *et al.* 2011, Gupta and Wang 2012). We use *sequential* to avoid confusion between online scheduling and web-based self-scheduling.

We posit that schedulers have an accommodation problem, not an optimization problem. That is, the scheduling of a patient calling to request a reservation (who is typically on the telephone, seeking an accommodation in the provider's appointment book) is done without full knowledge of the remaining number or type of patients who will eventually be scheduled within that session. At its core, the problem in such an environment is that assigning "any available slot" to "any patient" who prefers it may result in crowded sessions that incur provider overtime and patient delays.

Human schedulers are deft at managing the reservations of diverse customers who differ disparately by clinical needs and by consultation times. For example, Huang and Verduzco (2015) describe a women's clinic that classifies eight distinct customer types by consultation duration means and variances where new obstetric patients require three-times longer consultations than returning obstetric patients; similarly, the

978-1-5090-4486-3/16/$31.00 ©2016 IEEE

standard deviation of postpartum consultation times is nearly three times that of birth control management patients.

Under such heterogeneity of service times, the patient type sequence impacts an outpatient clinic session's operating characteristics dramatically (Klassen and Rohleder 1996, Cayirli *et al.* 2006). This is due to the accumulation of variation as a session evolves; templates that sort customers by the smallest-variance-first rule accumulate variation in customer finish times more slowly than other sequences, yielding minimum patient waiting and provider overtime (Wang 1999, Millhiser *et al.* 2012, Mak *et al.* 2014b). If the customer types seen on a given day are identified before their appointments are scheduled—known as *offline* scheduling—then their sequence and schedule can be optimized to minimize customer waiting and provider idle and overtime (Oh *et al.* 2013, Berg *et al.* 2014, Mak *et al.* 2014a, 2014b).

There are three barriers to the forgoing research making an impact on practice, however. First is that clinic managers are interested in balancing multiple conflicting operational and financial performance objectives simultaneously. For instance, the patient encounter should not be hurried for quality of patient care and clinical outcomes; waits should be controlled and fair; the clinic session should not run into overtime too long or too often; the clinic should serve as many people as possible to provide access and drive revenues; providers should not find themselves with chronic wasted idle times. Gupta and Wang (2012) acknowledge that mathematical models that optimize all criteria become intractable, and indeed we are aware of no model that optimizes these criteria for multiple patient types as satisfactorily as a human scheduler might. The motivation for our proposed decision support tool is to acknowledge the ability of human schedulers, and to give them new information to enable more effective balancing of these tradeoffs.

The second barrier is that the literature on optimal appointment template design is based on cost-based objective functions where costs increase linearly in the expected delays. Economists have estimated the cost of waiting per ambulatory care visit (Ray *et al.* 2015a) and by demographic (Ray *et al*. 2015b) in the United States. Similarly, Robinson and Chen (2010) suggested a novel way to estimate waiting costs per unit time, but empirical studies repeatedly show that customers are insensitive to short waits and grow increasingly sensitive to long waits, suggesting waiting costs grow exponentially in the delay (see Millhiser and Veral 2015 for references). Operational metrics such as the probability of delays and the probability of overtime capture nonlinear delay disutilities, alleviate the need to value the doctor's time (vis-à-vis the patient's) and as our informal interviews of ambulatory care managers in two large hospital systems in the New York Metropolitan Area suggest, are more intuitive for decision makers. Rather than scheduling to trade off difficult-to-estimate costs, scheduling to achieve probabilistic delay targets allows a clinic to space customer arrivals so that each customer enjoys about the same fair chance of waiting as every other customer (Turkcan *et al.* 2011, Huang *et al.* 2012, Sim and Qi 2012, Millhiser *et al.* 2012, Millhiser and Veral 2015). Scheduling with delay targets is more easily explained to the broad array of stakeholders including customers, schedulers, administrators, and clinicians. Furthermore, delay targets are akin to the widely-accepted service level agreements in other service industries (Millhiser and Veral 2015).

A third barrier is the reality that a small number of vendors dominate the electronic health record market (McCormack 2014). For example, consider Epic Systems which has the greatest market share (Gregg 2014). Given that Epic has an appointment-scheduling feature, and given how the dominance of Epic can slow IT innovation (Koppel and Lehmann 2015), the reality is that the foregoing scheduling algorithms will not replace the human scheduler who is increasingly supported by large, outsourced IT system. Even web-based self-scheduling services such as Zocdoc.com will not replace the ubiquitous human scheduler in the foreseeable future (Francis 2014).

To circumvent these barriers, we propose a decision support tool that informs the scheduler—at the moment a new customer calls to request an appointment—how that caller's service type affects the following session operating characterizes for every available slot within every day within the booking horizon: (a) the provider's probability of incurring overtime, (b) the calling patient's probability of incurring a service delay, and (c) the probability that the calling customer imposes a delay on existing customers in the appointment book. The decision support tool dynamically updates these probabilities after every appointment is added or subtracted (due to cancelations). The knowledge of (a), (b), and (c) informs the scheduler during each negotiation, and this information benefits both the clinic and the customer. In other words, rather than forcing a scheduling template upon a provider, the tool allows the scheduler to sequentially build the template, customer-by-customer, with the needs of the clinic and the needs of the customer in mind. It allows the scheduler to simultaneously manage the appointment delay and within-day waiting.

We submit that these three metrics are sufficient. We do not include the probability of provider idle time because idle time is corrected by scheduling more patients (and the scheduler is already able to trade-off the number of patients scheduled in a session and the conflicting goal of managed wait times). In other words, our decision support tool allows the scheduler to focus on the metrics that matter—the number of patients in the appointment book (i.e., revenue and access), the probability of overtime (i.e., operating cost), and the probability of patient delays (i.e., customer service) without distracting with additional information. Particularly noteworthy is the fact that this approach strips away the inherent subjectivity in traditional cost optimization approaches as delay costs must be subjectively and often arbitrarily assigned to both provider and patient delays.

The contribution of this work is new insights on sequential outpatient scheduling when (i) there are multiple customer types that differ by consultation time and duration, (ii) the clinic has collected empirical data on each type's historical service times and no-show probabilities, (iii) a clinic's strategic goal is to maximize customer choice in terms of allowing patients of all types to be booked at all times, (iv) a clinic's operational goals are to maximize fairness in customer waits while limiting session overtime and customer waiting, (v) and a human scheduler assigns customers times in the appointment book sequentially as they call (before the total session demand is known), and (vi) customers may be offered slots within a day or across multiple days. Furthermore, this paper is a first step at generalizing Millhiser and Veral (2015) in the case of multiple patient types.

2 LITERATURE REVIEW

The study of sequentially booking appointments is rooted in the so-called online scheduling theory; Sgall (1998) and Pruhs *et al.* (2004) summarize prior work in this area. In the healthcare application area, the dynamics and components of a sequential booking process are summarized in Gupta and Wang (2012).

For sequentially scheduling multiple patient types, it may be useful to review the performance of different rules toward this end. Walter (1973) demonstrated how clustering different patient types into homogeneous groups improves appointment system performance. Klassen and Rohleder (1996) studied customer types that differ by service time variances (with identical means), and confirm the efficacy of the smallest-variance-first rule (Pinedo and Wie 1986) for jointly minimizing customer waiting and provider idleness. Cayirli *et al.* (2006) expand Klassen and Rohleder's (1996) study across a wider range of environmental factors and customer types. Appointment slot durations are constant (not adjusted for the customer type), and so they find that placing returning patients first (i.e., padding the beginning of a session) yields less patient waiting and that placing new patients first (i.e., overbooking the beginning of the session) keeps provider idle time low. Cayirli *et al.* (2008) adjust the slot lengths to each type's mean consultation duration. A fixed service time coefficient of variation of 0.35 (not reported; verifiable in Cayirli 2004, pp. 107-108) implies that returning customers have smaller service time variance than new customers. They found, like Klassen and Rohleder (1996), that smaller-variance-first sequences minimize overall patient waiting and provider idle and overtime. They also show that if the value of the provider's time relative to the customers' time is sufficiently large, then the optimal sequence flips to new-customers-first without slot adjustment, i.e., starting the session with overbooking.

The foregoing studies fix the customer sequence prior to the booking period. In the present paper our sequential scheduling approach relaxes this requirement in favor of assigning customers to slots as they call in. In this case,“the best slot” is dynamically identified after every booking, but without considering future bookings.

The following five related papers model sequential scheduling under the assumption that customers are differentiated in their propensity to not show and are identical otherwise. Muthuraman and Lawley (2008) introduce sequential scheduling in a multiple-block scheduling system where customer assignment optimizes profit (rewards for each customer less costs of expected waiting and expected provider idle and overtime). The model was subsequently extended with other algorithms (Zeng *et al.* 2010), general service time distributions (Chakraborty *et al.* 2010), and dynamic programming models (Lin *et al.* 2011). Turkcan *et al.* (2011) employ an alternative objective that maximizes revenue subject to fairness constraints. They accept as many reservations as possible and assign every caller the appointment that yields the most even waiting experience of all patients scheduled so far. They stop accepting reservations when one of the following metrics exceeds a threshold: expected provider overtime, customer waits, or system congestion. Our metrics are similar, but we differ in that we classify patients by clinical needs (i.e., patient types/service time distributions) rather than no-show probabilities.

Zacharias and Pinedo (2014) extend the foregoing sequential scheduling models by differentiating customer types by their cost of waiting per unit time and their no-show probabilities. A heuristic identifies slots for overbooking to minimize expected customer waits and provider idle and overtime.

Our decision support tool is rooted in two assumptions about customer service—providers want to deliver fair within-day waiting and want to manage the appointment delay across days (say, such metrics as “time until first appointment,” “3rd-day out,” etc.). Toward the first end, Millhiser *et al.* (2012) assessed the fairness of common scheduling rules when customer types differ by mean and standard deviation of service durations. Introducing delay targets as a metric of operational performance, they show that common rules in the literature for scheduling multiple patient types lead to unfair waiting that accumulates at the end of a session, even when slot lengths are adjusted for customer type. As mentioned earlier, Turkcan *et al.* (2011) consider fair patient waits. So do Sim and Qi (2012) who propose assigning arrival times to a single class of customers so that a so-called “tardiness aware punctuality measure” keeps each customer’s Conditional Value at Risk below a threshold. In a different direction, Millhiser and Veral (2015) provide a simulation methodology that allows schedulers to build templates that offer each patient the same chance p of waiting at most t time units.

Toward the second end (managing appointment delays), several authors have considered the day (of multiple available days) as the decision variable. Patrick *et al.* (2008) propose a model for directing calling patients to the day that gives the fairest appointment delay for all. In a different direction, Liu *et al.* (2010) model this day choice to maximize a “net reward” when no-show rates are correlated with appointment delays. While there is intense pressure on schedulers to accommodate patients as soon as possible to minimize appointment delays, our tool allows them to see the operational impact of such decisions for the first time.

3 ANALYTICAL METHODOLOGY

We now describe the analytical framework and simulation methodology that underlie the decision support tool; our notation unifies that of Kaandorp and Koole (2007) and Liu *et al.* (2010). An outpatient clinic uses a separate appointment book for each physician. The scheduling horizon is of length D and at any moment appointment requests are being booked on days $d = 1, 2, 3, \ldots, D$ into the future. We assume there is a single session on day d that consists of T_d slots of duration Δ (Δ is set in most scheduling software, and thus the same every day). For example, a 4-hour session 3 days from now could be $T_3 = 48$ slots of length $\Delta = 5$ minutes. There are J patient types, each with a unique service time distribution with mean duration μ_j and standard deviation σ_j. The service time distributions are general and may be continuously or discreetly defined. A patient j’s type is further differentiated by the probability ρ_j of not showing, $j = 1, \ldots, J$. (Note that while the decision support tool we shall demonstrate at the conference uses the special case $\rho_1 = \cdots = \rho_J$, this limitation can be easily accommodated by the simulation model.)

On day d of the booking horizon, the vector $x_d = (x_{d1}, x_{d2}, \ldots, x_{dT_d})$ describes the current reservations in the appointment book where $x_{dt} = 0$ if there are no reservation on day d, slot t, and $x_{dt} = j$ if the reservation in slot t is a type-j patient. The sequence of vectors x_1, ..., x_D describes the current state of the appointment book across the scheduling horizon. In summary, the pair (d, t) denotes slot t on day d, and x_{dt} denotes the type of patient assigned to that slot, if any. The notation is summarized subsequently.

- D = the number of days in the booking horizon.
- $d = 1, 2, \ldots, D$ indexes the days within the booking horizon.
- Δ = length of a scheduling slot, in minutes.
- T_d = number of slots of length Δ available for bookings on day d.
- (d,t) = slot t on day d.
- J = number of customer types.
- $j \in \{1, 2, \ldots, J\}$ indexes the customer type.
- μ_j = mean service time of customer type j.
- σ_j = service time standard deviation, customer type j.
- ρ_j = probability that customer type j will not show with $0 \leq \rho_j < 1$.
- $x_{dt} \in \{0, 1, 2, \ldots, J\}$ = type of customer assigned current to slot (d,t) with $x_{dt} = 0$ if vacant.
- x_d = vector of assignments $x_d = (x_{d1}, x_{d2}, \ldots, x_{dT_d})$ on day d.
- $x_1, \ldots, x_D$ = current state of the appointment book.

This notation does not allow the double-booking of slots; we posit this is reasonable when the objective is fair and consistent wait times that achieve some targeted level of service. Numerical experiments suggest that slots as short as $\Delta = 5$ minutes are sufficient to achieve realistic delay targets.

Three events update the state of the appointment book x_1, ..., x_D. First, upon the end of the day, the index advances such that $x_d = x_{d+1}$ for $d = 1, \ldots, D-1$ and $x_D = (0, \ldots, 0)$. Second, x_{dt} increments from $x_{dt} = 0$ to $x_{dt} = j$ the moment a type-j patient books an appointment in slot (d,t). Finally, if the patient scheduled in slot (d,t) calls to cancel, then we set $x_{dt} = 0$.

Upon any appointment book state change, the probabilities of patient delays and provider overtime are reassessed. If arrivals were equally-spaced with no no-shows (i.e., $\rho_1 = \cdots = \rho_J = 0$), and service times Markovian and *i.i.d.*, then Jansson (1966) derives the necessary patient delay probabilities. During the booking period, however, the appointments in a partially-filled reservation book will not be equally-spaced, will have a propensity to not show, and will experience service times from non-identical and non-exponential distributions. For such generalizations, Millhiser and Valenti (2012) derive the probability distribution of each patient's finish time; from the cumulative distribution function (CDF) of any patient's finish time, one can compute the probability that the next scheduled patient will be delayed, say, 20 minutes. Similarly, the probability that session overtime will not exceed, say, 30 minutes, can be derived from the finish-time CDF of the last scheduled patient. They show that an arbitrary patient's finish-time CDF is a conditional sum of convolutions, namely, convolutions of service time distributions in a provider's "busy period" (the time of continuous service between successive idle times), conditioned on when each busy period started. With this stochastic model, one may assess the probability of delay for each patient in any appointment book on any day of the booking period.

4 IMPLEMETATION VIA SIMULATION

The aforementioned analytical model, due to its need for numerical integration of the convoluted finish-time distributions, faces implementation challenges. Numerical integration allows performance analysis of a *given* template, but does not lead to the development of templates that adhere to the operational performance goals. In other words, numerical integration calculates overtime and delay probabilities given a *fixed* schedule, whereas finding a schedule that achieves desired delay probabilities requires tedious numerical methods that cannot be run in real time. This problem was averted in previous research via an inverse simulation modeling approach that sequentially assigns patients, adjusting the slots so that each successive patient has a predetermined chance of a wait exceeding a targeted duration (Millhiser and Veral

2015). This leads to a set of arrival slots for homogenous patients. The problem was also averted in a prior simulation study by assigning arrival times based on individuals' (non-convoluted) service time distributions, rather than finish time distributions (Huang *et al.* 2012).

Millhiser and Veral (2015) adjust the arrival slots to the multiples of 5, 10, or 15-minute slots (e.g., $\Delta = 5$, 10, or 15 minutes). The result is an "ideal" template based on the assumption that the schedulers will be able to fill all slots with the designated homogenous class of patients. Figure 1 depicts one such template, where rounded slots are adjusted so that all patients have 80% or higher chance of waiting for no more than 20 minutes. The bars in Figure 1 depict the slot length between successive arrivals, and the superimposed line graph shows the associated probability that each scheduled patient will wait at most 20 minutes.

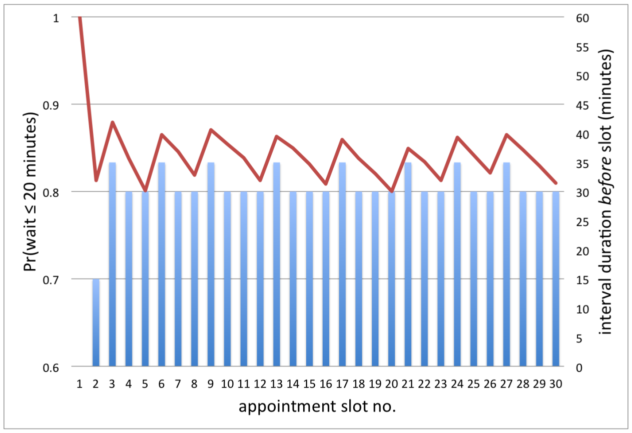


Figure 1: Operational performance of a template when service times are gamma-distributed with mean 30 minutes and coefficient of variation 0.2. The left axis denotes the probability of 20-minute delays indicated by the solid line (source: Millhiser & Veral 2015, Figure 4b.).

Outpatient clinics that schedule appointments via telephone, however, serve numerous types of patient classes with differing consult lengths, and the patient preferences preclude providers' ability to populate each session slot sequentially, without gaps between successive appointment requests.

The proposed decision support tool takes advantage of Millhiser and Veral's (2015) simulation-based probability analysis, and extends it to partially-booked sessions of heterogenous patient types. We assume that the moment a patient calls to request a booking, the scheduler and patient have designated a mutually agreeable day $d \in \{1,\ldots,D\}$ (e.g., the provider is in, the patient has available time, and the template is partially filled with unreserved slots) and the scheduler can ascertain the calling patient's type (e.g., service time characteristics of the type of consultation required). Equipped with this information, the simulator checks the start time of every available slot (i.e., every slot (d,t) with $x_{dt} = 0$) and simulates R replicates of the entire session, with the patient calling for the reservation request placed in each of the available slots.

For each patient type j, we assume that there is a history of n_j historical service times, and so for every simulation run we draw each booked patient's random service time using the well-known bootstrapping method. (Alternatively, fitted historical service time distributions that pertain to the corresponding patient type can be used for random number generation.)

4.1 The probability of the next calling patient's delay

We assume that the provider has specified a target probability of delay denoted $[P,M]$ where P is the desired probability that a patient waits at most M minutes from the scheduled time of arrival until the time the patient is actually seen by the provider. First, the simulation computes the probability that the delay does not exceed M minutes for every unbooked slot into when the next calling patient may be placed. Let

$A_d = \{t : x_{dt} = 0\}$ be the set of available (unreserved) slots on day d. On a given day d, we specifically denote the probability $p_t = \Pr(\text{delay} \leq M)$ for all $t \in A_d$. Note that p_t is a function of the booked patients in the appointment book, not the type of the calling patient, and so p_t can be computed immediately after any change in appointment book state.

4.2 The probability of existing reservations' delays

When the next calling patient's type is revealed, the probability of delays for existing reservations may be computed via simulation. We let $B_d = \{t : x_{dt} > 0\}$ be the set of slots on day d that are booked with a prior reservation. The probability of delay of an existing reservation depends on which available slot the next calling patient is placed. On day d, let $q_{s|t}$ be the probability that the existing reservation in slot $s \in B_d$ be delayed at most M minutes given that the new calling patient is assigned to slot $t \in A_d$.

We exemplify in Figure 2 with a simple one-hour session on some future day d consisting of 12 slots of $\Delta = 5$ minutes. The session ends at 9:00 at which time overtime may be incurred. There are five reservations in he appointment book; they called in the order A, B, C, D, and E, and made reservations for the slots indicated in column (a). That is, column (a) depicts the initial session profile and available slots prior to the arrival of the 6th reservation request, i.e., $A_d = \{1, 5, 7, 8, 9, 11, 12\}$ and $B_d = \{2, 3, 4, 6, 10\}$. Column (b) gives the delay probabilities p_t for each available slot $t \in A_d$. Column (c) gives the delay probabilities $q_{s|t}$ for each booked slot $s \in B_d$ given every possible "next" reservation in slot $t \in A_d$. These probabilities are extended to the session end with p_{se} denoting that the probability that the existing bookings incur overtime of at most M_{se} minutes, and the related probabilities of overtime $q_{se|t}$ if the next caller is assigned to slot $t \in A_d$.

			Pr(delay $\leq M$ minutes)	
Interval	Time	(a) Reservation	(b) New Patient	(c) Existing Reservations
1	8:00		$p_1 = 1$	
2	8:05	Patient E		$q_{2\|1}, q_{2\|5}, q_{2\|7}, q_{2\|8}, q_{2\|9}, q_{2\|11}, q_{2\|12}$
3	8:10	Patient A		$q_{3\|1}, q_{3\|5}, q_{3\|7}, q_{3\|8}, q_{3\|9}, q_{3\|11}, q_{3\|12}$
4	8:15	Patient D		$q_{4\|1}, q_{4\|5}, q_{4\|7}, q_{4\|8}, q_{4\|9}, q_{4\|11}, q_{4\|12}$
5	8:20		$p_5 < 1$	
6	8:25	Patient B		$q_{6\|1}, q_{6\|5}, q_{6\|7}, q_{6\|8}, q_{6\|9}, q_{6\|11}, q_{6\|12}$
7	8:30		p_7	
8	8:35		p_8	
9	8:40		p_9	
10	8:45	Patient C		$q_{10\|1}, q_{10\|5}, q_{10\|7}, q_{10\|8}, q_{10\|9}, q_{10\|11}, q_{10\|12}$
11	8:50		p_{11}	
12	8:55		p_{12}	
End	9:00	Session End	p_{se}	$q_{se\|1}, q_{se\|5}, q_{se\|7}, q_{se\|8}, q_{se\|9}, q_{se\|11}, q_{se\|12}$

Figure 2: A sample template with 5 existing bookings.

Note that $p_1 = q_{2|5} = q_{2|7} = q_{2|8} = q_{2|9} = q_{2|11} = q_{2|12} = 1$ as the first patient experiences no delay. However $q_{2|1} < 1$ as assigning the calling patient to slot 1 would delay Patient E (slot 2) with positive probability. If the calling patient reserves slot 5, then $p_5 < 1$ as there would be considerable likelihood that three consecutive consultations staring at 8:05, 8:10, and 8:15 would create an excessive delay for slot 5; furthermore, if slot 5 is booked by the 6th patient ("F"), then Patient B in the 6th slot would have even a higher probability that his/her wait would exceed M minutes.

A user-friendly presentation of the resulting probabilities is implemented through color-coding to indicate those probabilities that are between P and $P - 10\%$ (yellow); and less than $P - 10\%$ (red). This information allows the scheduler to deny the "red" slots and use discretion about populating the "yellow" slots while negotiating with the patient.

4.3 Summary of simulation procedure

Summarizing the above discussion, we use the following simulation routine for reporting the operational characteristics of the current appointment book.

1. Suppose a type-j patient calls to request a reservation, $j \in \{1,\ldots,J\}$.
2. For every day d and every available slot $\hat{t} \in A_d$, repeat the following steps.
   (a) Initialize a counter $C_{dt} = 0$ for every slot (d,t), $t = 1,\ldots,T_d$.
   (b) Temporarily force the next caller to be assigned slot $\hat{t}$ and update the set of booked slots $\hat{B}_d = B_d \cup \{\hat{t}\}$.
   (c) Assume that patients and providers arrive punctually such that the first scheduled arrival of each day starts on time.
   (d) For every booked slot (d,t) with $t \in \hat{B}_d$ with an assigned patient of type $x_{dt} > 0$, draw a binomial random variable with probability $\rho_{x_{dt}}$ of not showing, and if the patient shows, then draw a random number from the service time distribution of a type-x_{dt} patient.
   (e) Compute the delay for every slot (d,t) with $x_{dt} > 0$ (and the patient showing), given each prior patient's finish time. For every slot (d,t) where the delay does not exceed M minutes, increment $C_{dt} = C_{dt} + 1$.
   (f) Repeat steps 2(c) to 2(e) for R simulation replicates.
   (g) For every slot (d,t) assigned patient type $x_{dt} > 0$, compute the probability C_{dt}/R. Let $p_{\hat{t}} = C_{d\hat{t}}/R$ and $q_{t|\hat{t}} = C_{dt}/R$ for all $t \in \hat{B}_d \setminus \{\hat{t}\}$.
3. At the scheduler's discretion, assign the calling patient of type j to a slot in set A_d using the information in step 2(g) as a guide.
4. Go to step 1.

The decision support tool reports the information given in step 2(g) to inform the decision in step 3. We coded the simulation in Java; run times to executive the above steps using 1 million simulation replications on 8-hr sessions has not exceeded 5 seconds on a 2.7 GHz Intel Core i7 processor; the time to rerun the simulation after every patient booking is not a hindrance in practice.

5 EXAMPLE AND DISCUSSION

We give an illustrative example of the foregoing methodology using a data set reported by Huang and Verduzco (2015): 8 patient classes whose mean and standard deviation of consultation times at a women's health clinic are summarized in Figure 3. Note how expected service times conveniently fit into traditional 15- and 30-minute slots with minimal rounding. Suppose that the provider stated performance targets of 90% of sessions experiencing less than 30 minutes overtime, and 80% of patients waiting less than 20 minutes to be seen. For demonstration purposes, we assume a 5% no-show rate for all patient types. Consider a 4-hour session from 8:00 AM to 12:00 PM on Wednesday, March 15, 2017, and suppose at this moment (prior to March 15) there are 5 existing reservations shown in Figure 4 and a Level 2 patient is calling to request a March 15 appointment. Upon identifying the patient type, the decision support tool simulates the session and reports the probabilities shown in Figure 5(a). The following interpretations of Figure 5(a) may be useful.

1. The left column in Figure 5 gives times in 5-minute increments (we assume the scheduling software can accommodate such slots).

Code	Patient Type	Mean	Std. Dev.	N
A	Birth Control Mgmt	14.3	3.7	20
B	Follow-up	16.1	11.3	22
C	Level 1	25.0	7.4	39
D	Level 2	37.1	10.3	32
E	New Obstetric	37.1	10.9	38
F	Postpartum	26.1	12.1	17
G	Physical Exam	33.7	11.8	50
H	Routine Obstetric	13.0	5.7	207

Figure 3: Consultation time distributions in minutes for 8 patient types (source: Huang & Verduzco 2015).

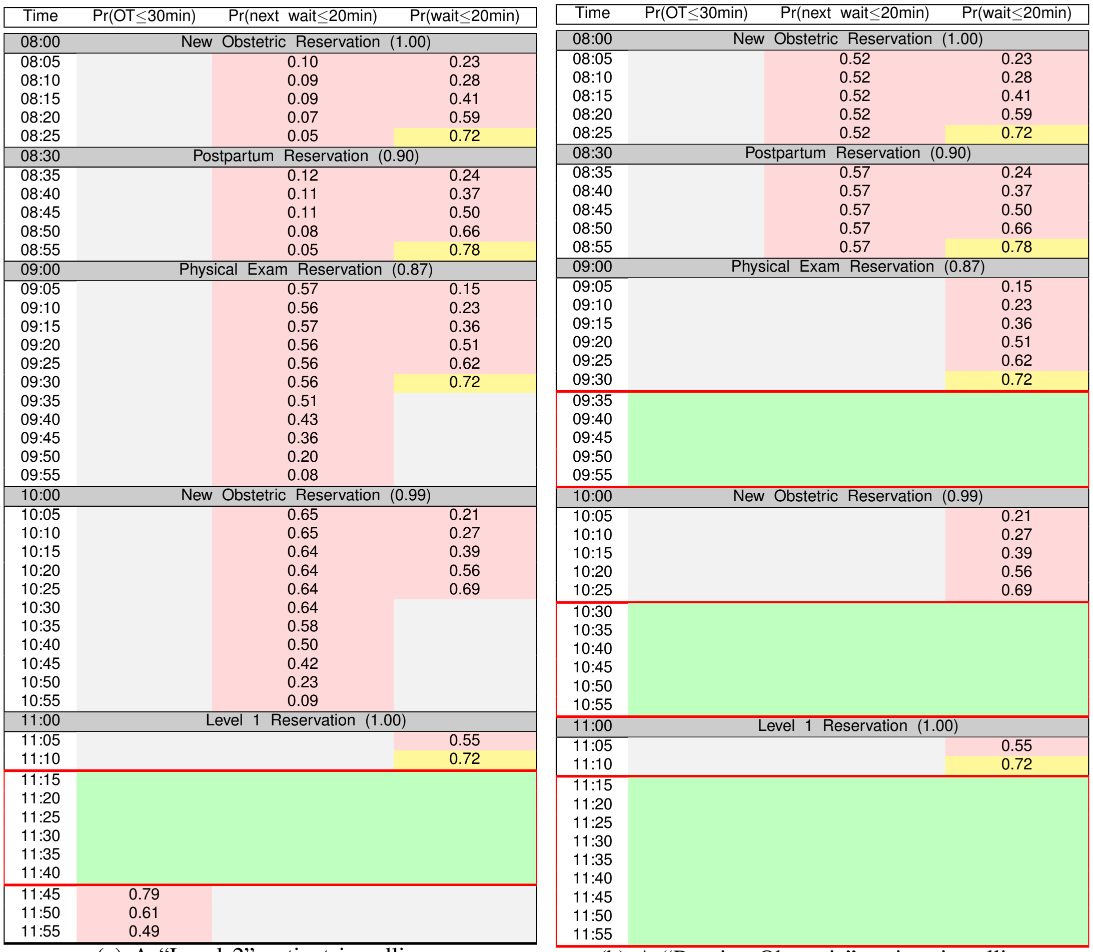
8:00 AM	8:30 AM	9:00 AM	10:00 AM	11:00 AM
New Obstetric	Postpartum	Physical Exam	New Obstetric	Level 1

Figure 4: Five existing reservations used in numerical example.

2. Any 5-minute slot with an existing reservation is "greyed-out" and displays the patient type and the probability that this patient incurs at most 20-minutes' wait given the existing 5 reservations. This prevents the scheduler from double-booking any one 5-minute slot. For example, the scheduled Postpartum patient at 8:30 presently has a 90% chance of waiting less than 20 minutes.
3. Empty cells are those with probability greater than 0.9 for the OT column, and 0.8 for the two patient wait columns (hidden to simplify the output).
4. The second column ("Pr(OT $\leq$ 30 min)") gives for each slot the probability that the session incurs at most 30 minutes' overtime if the calling Level 2 patient is assigned to that slot. Reservations at 11:45 or later violate the 90% target and are hence colored light red. On this basis only, it is advised that the Level 2 patient be assigned 11:40 or earlier.
5. The third column ("Pr(next wait $\leq$ 20 min)") indicates for every available slot the chance that the existing reservation immediately following the slot waits at most 20 minutes if the Level 2 patient were assigned that slot. For example, if the calling Level 2 patient is given 9:30, then 10:00 Obstetric patient will have 56% chance of waiting at most 20 minutes. On this basis, inserting the Level 2 patient prior to 11:05 will cause an existing reservation a delay that does not meet the performance target (and such cells are hence coded red).
6. The fourth column ("Pr(wait $\leq$ 20 min)") indicates for each slot the probability that the calling Level 2 patient will wait at most 20 minutes if assigned to that slot. Slots that miss the target by no more than 10% (i.e., 70% to 80%) are shaded yellow; slots that miss the target by more than 10% are shaded light red.
7. The gray shaded cells show the feasible slots with respect to each constraint independently. Only slots 11:15 to 11:40 satisfy all three criteria, as indicated with the green cells with red outlining, prompting the scheduler to assign the calling Level 2 patient to one of these 6 slots if possible.
8. While the decision support tool does not allow double-bookings on top of an existing reservation, two bookings 5 minutes apart are possible, and the decision support tool indicates that in this case one would not do this in the interest of the performance goals.

Figure 5(a) is meant to exemplify that value of the decision support tool. Most likely, in practice, such a clinic would work with 15- and 30-minute slots. A scheduler in a traditional scheduling system would likely offer 9:30, 10:30, and 11:30 as viable times. However, the decision support tool makes it clear that the patient should be directed to the 11:30 slot or a different day. Moreover, the software gives the scheduler full information—that any slot from 11:15 to 11:40 (not only 11:30) allow all three performance criteria to be met.

Another benefit of the software is that as it dynamically recalculates the probabilities with each subsequent reservation, and it also dynamically recalculates the probabilities for every patient type. For



Time	Pr(OT≤30min)	Pr(next wait≤20min)	Pr(wait≤20min)
08:00	New Obstetric Reservation (1.00)		
08:05		0.10	0.23
08:10		0.09	0.28
08:15		0.09	0.41
08:20		0.07	0.59
08:25		0.05	0.72
08:30	Postpartum Reservation (0.90)		
08:35		0.12	0.24
08:40		0.11	0.37
08:45		0.11	0.50
08:50		0.08	0.66
08:55		0.05	0.78
09:00	Physical Exam Reservation (0.87)		
09:05		0.57	0.15
09:10		0.56	0.23
09:15		0.57	0.36
09:20		0.56	0.51
09:25		0.56	0.62
09:30		0.56	0.72
09:35		0.51	
09:40		0.43	
09:45		0.36	
09:50		0.20	
09:55		0.08	
10:00	New Obstetric Reservation (0.99)		
10:05		0.65	0.21
10:10		0.65	0.27
10:15		0.64	0.39
10:20		0.64	0.56
10:25		0.64	0.69
10:30		0.64	
10:35		0.58	
10:40		0.50	
10:45		0.42	
10:50		0.23	
10:55		0.09	
11:00	Level 1 Reservation (1.00)		
11:05			0.55
11:10			0.72
11:15	🟩	🟩	🟩
11:20	🟩	🟩	🟩
11:25	🟩	🟩	🟩
11:30	🟩	🟩	🟩
11:35	🟩	🟩	🟩
11:40	🟩	🟩	🟩
11:45	0.79		
11:50	0.61		
11:55	0.49		

(a) A “Level 2” patient is calling.

Time	Pr(OT≤30min)	Pr(next wait≤20min)	Pr(wait≤20min)
08:00	New Obstetric Reservation (1.00)		
08:05		0.52	0.23
08:10		0.52	0.28
08:15		0.52	0.41
08:20		0.52	0.59
08:25		0.52	0.72
08:30	Postpartum Reservation (0.90)		
08:35		0.57	0.24
08:40		0.57	0.37
08:45		0.57	0.50
08:50		0.57	0.66
08:55		0.57	0.78
09:00	Physical Exam Reservation (0.87)		
09:05			0.15
09:10			0.23
09:15			0.36
09:20			0.51
09:25			0.62
09:30			0.72
09:35	🟩	🟩	🟩
09:40	🟩	🟩	🟩
09:45	🟩	🟩	🟩
09:50	🟩	🟩	🟩
09:55	🟩	🟩	🟩
10:00	New Obstetric Reservation (0.99)		
10:05			0.21
10:10			0.27
10:15			0.39
10:20			0.56
10:25			0.69
10:30	🟩	🟩	🟩
10:35	🟩	🟩	🟩
10:40	🟩	🟩	🟩
10:45	🟩	🟩	🟩
10:50	🟩	🟩	🟩
10:55	🟩	🟩	🟩
11:00	Level 1 Reservation (1.00)		
11:05			0.55
11:10			0.72
11:15	🟩	🟩	🟩
11:20	🟩	🟩	🟩
11:25	🟩	🟩	🟩
11:30	🟩	🟩	🟩
11:35	🟩	🟩	🟩
11:40	🟩	🟩	🟩
11:45	🟩	🟩	🟩
11:50	🟩	🟩	🟩
11:55	🟩	🟩	🟩

(b) A “Routine Obstetric” patient is calling.

Figure 5: The decision support tool when there are five existing bookings and the patient calling to request an appointment is of type (a) “Level 2” and (b) “Routine Obstetric.”

instance, rather than a Level 2, suppose that a Routine Obstetric patient is on the phone requesting a March 15 appointment, given the same existing 5 reservations as before. The decision support tool would give different probabilities shown in Figure 5(b). The second column now indicates that a Routine Obstetric patient (with short service times of mean 13.0 and standard deviation 5.7 minutes) does not jeopardize the 30-minute overtime target, even if assigned 11:55. The third column in Figure 5(b) shows 22 more slots where the Routine Obstetric patient has an 80% or better chance of not delaying an existing reservation 20 minutes. In other words, it is possible to “squeeze in” a Routine Obstetric patient anywhere between the 3rd and 4th reservations or between the 4th and 5th reservations, whereas a Level 2 patient cannot. Finally, the fourth columns of Figures 5(a) and 5(b) are identical, as one expects given that the calling patient’s delay is due to the existing reservation types.

While the decision support tool aids the scheduler in not violating service level targets, it does not attempt to optimize the schedule with respect to throughput or utilization. Most notably, it does not prevent the scheduler from inserting unnecessarily long gaps between successive appointments. To exemplify, in Figure 5(a), the “recommended” slots for a calling Level 2 patient are 11:15 through 11:40. If the scheduler places the patient in the 11:20 slot, then the 11:15 slot will not accommodate any future patient (as any placement would violate the waiting time target for the 11:20 slot), resulting in a 5-minute gap that is “wasted”. Likewise, if this future calling patient were scheduled in the 11:25 slot, the 11:15 and 11:20

slots would not be adequate for longer consult types (e.g., codes C-G in Figure 3). To ameliorate the inefficiencies that may arise from this dynamic, schedulers should be urged to place patients in the first (i.e., 11:15) or the last (i.e., 11:40) slot in any block of multiple contiguous available slots.

Even though this is a shortcoming from scheduling theory perspective, it is more reflective of operational reality in real-time appointment scheduling. Future work in this field may incorporate optimization efforts for static scheduling environments where demand is known in advance. Examples may include elective surgeries that are scheduled for the relatively distant future or inpatient radiology scheduling where most physicians request service in the early morning, and the requests may be batched prior to scheduling.

6 SUMMARY

At the time of this writing the authors are preparing to pilot the decision support tool in July-August 2016 at a major New York City healthcare network. The presentation given at the December 2016 Winter Simulation Conference will summarize the foregoing research and simulation methodology. We will give a demonstration of the decision support tool and share results and insights from the pilot study (if available).

In summary, we have presented a simulation methodology that we believe is the first of its kind to allow schedulers in ambulatory care services to dynamically monitor the affect of booking heterogeneous calling patients on a scheduling template's operational performance. Based on prior research, we have expressed those performance measures as probabilities of delay and overtime which we believe to be a natural/intuitive way to express risk and uncertainty.

REFERENCES

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.

Cayirli, T. 2004. *Ambulatory care performance: A simulation study of the role of appointment scheduling rules, patient classification and environmental factors*. PhD Dissertation, Baruch College, The City University of New York.

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

Francis, J. 2014. "Two-in-three patients will book medical appointments online in five year, Accenture forecasts." Newsroom [blog], December 9, 2014, http://bit.ly/29Hcma7 (retrieved 14 July 2016).

Gregg, H. 2014. "50 things to know about Epic, Cerner, MEDITECH, McKesson, Athenahealth and other major EHR vendors." *Becker's Health IT & CIO Review*, July 14, 2014, http://bit.ly/29MzUxb (retrieved 14 July 2016).

Gupta, D., and Wang, W.Y. 2012. "Patient appointments in ambulatory care." Chapter 4 in *Handbook of Healthcare System Scheduling* (pp. 65-104), R. Hall (ed.), Springer US.

Herrler, R., Heine, C., and Kluegl, F. 2002. "Appointment scheduling among agents: A case study in designing suitable interaction protocols." AMCIS 2002 Proceedings, 199.

Huang, Y. L., Hancock, W. M., and Herrin, G. D. 2012. "An alternative outpatient scheduling system: Improving the outpatient experience." *IIE Transactions on Healthcare Systems Engineering* 2(2): 97-111.

Huang, Y., and Verduzco, S. 2015. "Appointment template redesign in a women's health clinic using clinical constraints to improve service quality and efficiency." *Applied Clinical Informatics* 6(2): 271-287.

Kaandorp, G. C., and Koole, G. 2007. "Optimal outpatient appointment scheduling." *Health Care Management Science* 10(3): 217-229.

Klassen, K.J., and Rohleder, T.R. 1996. "Scheduling outpatient appointments in a dynamic environment." *Journal of Operations Management* 14(2): 83-101.

Koppel, R., and Lehmann, C.U. 2015. "Implications of an emerging EHR monoculture for hospitals and healthcare systems." *Journal of the American Medical Informatics Association* 22(2): 465-471.

Liu, N., Ziya, S., and Kulkarni, V.G. 2010. "Dynamic scheduling of outpatient appointments under patient no-shows and cancellations." *Manufacturing & Service Operations Management* 12(2): 347-364.

Mak, H.Y., Rong, Y., and Zhang, J. 2014a. “Sequencing appointments for service systems using inventory approximations.” *Manufacturing & Service Operations Management* 16(2): 251-262.

Mak, H.Y., Rong, Y., and Zhang, J. 2014b. “Appointment scheduling with limited distributional information.” *Management Science* 61(2): 316-334.

McCormack, M. 2014. “EHR meaningful use market share.” SoftwareAdvice.com, http://www.softwareadvice.com/resources/ehr-meaningful-use-market-share (retrieved 14 July 2016).

Millhiser, W.P., Valenti, B.C. 2012. “Delay distributions in appointment systems with generally & non-identically distributed service times & no-shows.” Available at SSRN: http://ssrn.com/abstract=2045074.

Millhiser, W. P., Veral, E. A., and Valenti, B. C. 2012. “Assessing appointment systems’ operational performance with policy targets.” *IIE Transactions on Healthcare Systems Engineering* 2(4): 274-289.

Millhiser, W.P., and Veral, E.A. 2015. “Designing appointment system templates with operational performance targets.” *IIE Transactions on Healthcare Systems Engineering* 5(3): 125-146.

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.

Patrick, J., Puterman, M. L., and Queyranne, M. 2008. “Dynamic multipriority patient scheduling for a diagnostic resource.” *Operations Research* 56(6): 1507-1525.

Pinedo, M., and Wie, S. H. 1986. “Inequalities for stochastic flow shops and job shops.” *Applied Stochastic Models and Data Analysis* 2(1-2): 61-69.

Pruhs, K., Sgall, J., and Torng, E. 2004. “Online scheduling.” In *Handbook of Scheduling: Algorithms, Models, & Performance Analysis*, J.Y-T. Leung (ed.), Chapman & Hall/CRC: Boca Raton, Florida.

Ray, K. N., Chari, A. V., Engberg, J., Bertolet, M., and Mehrotra, A. 2015a. “Opportunity costs of ambulatory medical care in the United States.” *The American journal of Managed Care* 21(8): 567-574.

Ray, K. N., Chari, A. V., Engberg, J., Bertolet, M., and Mehrotra, A. 2015b. “Disparities in time spent seeking medical care in the United States.” *JAMA Internal Medicine* 175(12): 1983-1986.

Sgall, J. 1998. “On-line scheduling.” In *Online Algorithms* (pp. 196-231). Springer: Berlin.

Sim, M., and Qi, J. 2012. “Outpatient appointment sequencing and scheduling under uncertainty.” 23rd Annual Production & Operations Management Conference, Chicago, IL, April, 22, 2012.

Turkcan, A., Zeng, B., Muthuraman, K., and Lawley, M. 2011. “Sequential clinical scheduling with service criteria.” *European Journal of Operational Research* 214(3): 780-795.

Walter, S. D. 1973. “A comparison of appointment schedules in a hospital radiology department.” *British Journal of Preventive & Social Medicine* 27(3): 160-167.

Zacharias, C., and Pinedo, M. 2014. “Appointment scheduling with no-shows and overbooking.” *Production & Operations Management* 23(5): 788-801.

AUTHOR BIOGRAPHIES

WILLIAM P. MILLHISER is an Associate Professor of Operations Management with research interests in stochastic modeling and simulation modeling, especially in the optimization of queueing systems such as appointment scheduling systems. He teaches Operations Management in the MBA and BBA programs, and is the recipient of the 2013 Baruch College President’s Award for Distinguished Teaching. His email address is william.millhiser@baruch.cuny.edu.

Emre A. Veral is the Academic Director of the Baruch MBA Program in Health Care Administration and Professor of Operations Management in the Loomba Department of Management. His current research agenda focuses on Appointment System Design and operational effectiveness of Health Care Delivery Systems. Professor Veral also serves as Faculty Mentor/Consultant at the Field Center for Entrepreneurship at Baruch College. His email address is emre.veral@baruch.cuny.edu.