Appointment Scheduling at Outpatient Clinics Using Two-Stage Stochastic Programming Approach

SAMIRA FAZEL ANVARYAZDI\textsuperscript{1}, SARAVANAN VENKATAChALAM\textsuperscript{2}, AND RATNA BABU CHINNAM\textsuperscript{2}\
\textsuperscript{1}Olin Business School, Washington University in St. Louis, St. Louis, MO 63130, USA\
\textsuperscript{2}Department of Industrial and Systems Engineering, Wayne State University, Detroit, MI 48201, USA\
Corresponding author: Saravanan Venkatachalam (saravanan.v@wayne.edu)

ABSTRACT Clinics with a large volume of patients are often burdened with limited resources such as nurses and providers. Also, an efficient health system seeks short wait times for the patients to see the provider (indirect wait time) and within the clinic (direct wait time) during the day of the appointment. Additionally, the appointment duration, volume of patients, no-show behavior are uncertain. The direct and indirect wait times, stochastic parameters, rising treatment costs, and increased demand of patients motivate the need for efficient appointment schedules and clinic operations. In this paper, we develop a two-stage stochastic mixed-integer linear programming model (SMILP) integrated with a simulation model to generate a scheduling template for the providers to schedule individual patient appointments and resources. The model minimizes the expected wait times for the patients with a fair and equitable utilization of the resources. Computational experiments were conducted using a data-driven simulation model, and the results indicate that the proposed approach can significantly decrease patients’ direct and indirect wait times when compared to a deterministic indexing policy used for scheduling appointments.

INDEX TERMS Appointment scheduling, direct wait time, healthcare, indexing policy, indirect wait time, patient no-show, two-stage stochastic programming.

I. INTRODUCTION

A variety of studies have documented the substantial deficiencies in the quality of healthcare delivered across the United States ([1]–[5]). Attempts to reform the United States health care system in the 1980s and 1990s were spearheaded by the system’s inability to adequately provide access, ensure quality, and restrain costs, but these efforts had limited success. In the era of managed care, there are challenges in access, quality, and costs, and also medical professionals are increasingly dissatisfied [6]. According to the Centers for Medicare and Medicaid Services (CMS) [7], United States devotes 17.7% of GDP to health care (spending $11,172 per person in 2018, or $3.6 trillion total), yet life expectancy is slightly low, and nearly 28 million Americans were uninsured in 2018 [8]. While there are no comparable studies of the quality of care delivered in the hospital outpatient setting, pervasive deficits across the health system suggest similar problems likely exist. Especially, a large fraction of care delivered in this setting is ambulatory care for acute and chronic conditions where deficits in quality have been amply demonstrated [9]. In addition to the potential deficits in the quality of care in the hospital outpatient setting, CMS and others have also observed growth in the volume of services and costs for care delivered in this setting. In 2006, care provided to Medicare beneficiaries in the hospital outpatient setting accounted for 7% of total Medicare program spending [10], and overall spending nearly doubled between 1996 and 2006, reaching $31.6 billion [11]. Reported by CMS Office of the Actuary projects, total spending increased on average by 10.4% per year from 2003 to 2008 [11]. A proposed payment system for hospital outpatient services was accomplished in the year 2000 and the services paid under it show approximately 90% of spending on all hospital outpatient services. All these studies emphasize the importance of performance for outpatient services.

Developing an efficient appointment scheduling and management system considering a stochastic server is needed...
to overcome the following problems: the no-show behavior of patient arrival, patient/provider check-in delays, overbookings, long wait times, and poor provider/staff utilization. These are pervasive in outpatient clinics ([12]–[21]). On the other hand, appointment scheduling systems, which give patients flexibility in choosing their appointment time, not only lead to satisfied patients but also have outstanding effects on other patient flow metrics such as decreasing the no-show rate as well as patient appointment delays (time between patient desired time and assigned appointment time) and higher patient retention rates. All of these result in better reimbursement rates by payers for providers ([22], [23]).

Many studies have documented the no-show rate in medical practice. The study in [19] reported a 42% average no-show rate which ranges from 6% to 92% in outpatient clinics. Also, the study in [20] reports 13% to 24% no-show rates at endoscopy clinics for different service types, and [24] indicates post intervention no-show rates ranging from 28% to 45%. The results in [25] show the overall no-show rate at Obstetrics and Gynecology (OBGYN) clinics as 30.1%. They investigated the strong relationship between patients’ appointment delays and no-show cases in OBGYN specialty clinics, and similarly in psychotherapy appointments, a 21% no-show rate was reported by [21].

Another patient flow metric is patients’ appointment delays known as indirect wait time in the literature. A survey was conducted by [26] to determine the average indirect wait time for new patients to see a provider in 15 major and 15 mid-sized metropolitan areas in different specialty clinics. The survey was organized in 2004, 2009, 2014, and 2017, and the results show an increase in the indirect wait time in 2017 compared to other years. For instance in 2004, it was reported for 15 mid-sized metropolitan markets between 88,000 and 143,000 people including 1,414 medical offices in large metro markets and 494 medical offices in mid-sized metro markets. The report on indirect wait time for cardiology, dermatology, obstetrics and gynecology, orthopedic surgery and family medicine, and average indirect wait time of obstetrics and gynecology was studied in [26]. The average is about 25 days with a standard deviation of 10 days. In this study, we develop a framework to minimize direct and indirect wait times simultaneously while considering fair and equitable utilization of the resources in a clinic. A case study based on the operations in OBYGN clinic is also presented to corroborate the performance of the approach.

In this paper, we present an integrated approach of two-stage stochastic and simulation models to generate scheduling template while considering the sources of uncertainties such as no-show behavior, direct and indirect wait times, and fair and equitable assignments of new patients to the providers. Following are the major contributions of this paper: 1) a two-stage stochastic model to generate a weekly scheduling pattern for individual providers that gets updated at regular intervals based on the type and mix of services rendered; 2) a clinic simulation model to determine sequencing rules to maintain the direct wait times within a threshold; 3) extensive data-driven simulated computational experiments to corroborate the effectiveness of the proposed method with a first-call-first-serve policy used in the call center. The overall framework is depicted in Figure 1. This paper is organized as follows. Section II reviews the relevant literature and describes the problem. Section III discusses the development of our model formulation and solution scheme. Section IV presents the computational experiments and insights of the results based on data driven simulation study for an OBYGN clinic. Finally, Section V provides concluding remarks.

II. RELEVANT LITERATURE

In recent years, appointment scheduling in outpatient clinics has attracted much attention in health care delivery systems. The authors in [27] and [28] discussed a variety of methods in modeling, optimization, and future work in appointment scheduling. Designing an effective appointment scheduling system in outpatient clinics helps a smooth flow of patients while considering patients’ and physicians’ preferences, and matching supply and demand. The research in [29] developed a framework to find the possibility that patients may cancel or no-show at their time of appointments. The authors in [28] divided decision making in outpatient appointment scheduling into static and dynamic models. The work in [28] and [30] define static scheduling as the decisions about appointment times that are made prior to the start of the appointment session while in dynamic case, the schedule may be modified later depending on the state of the system. The work in [31] considered a static appointment scheduling where the number of patients is already known. The work in [29] developed a dynamic scheduling where appointments are assigned to patients depending on the clinic’s appointment schedule at the time of the patient’s call. In our work, we consider the dynamic model for our computational experiments, and the proposed two-stage model proposes a scheduling template for the call center scheduler.

Another classification on appointment scheduling is with respect to the wait times. The work in [29], [30], [32], [33], and [34] are among the most recent studies on appointment scheduling which consider wait time in model formulations. Other than wait times, the research in [34]
addresses crucial characteristics such as the randomness of service time, patient punctuality and no-show behavior. Similar to the research in [34], we minimize indirect and direct wait times in the model by determining appropriate schedule template for the providers. In another study by [33], a weekly schedule pattern in outpatient clinics for an OB/GYN specialty is determined while considering only direct wait time for the patients.

A two-stage stochastic integer programming approach is one of the most common methods in appointment scheduling ([20], [30], [32], [33] and [31]). The authors in [20] presented a two-stage stochastic mixed-integer program considering uncertainty in a system for optimizing booking and appointment times with the objective of maximizing expected profit for outpatient clinics. The research work in [32] formulated the appointment scheduling in two steps: a two-stage stochastic linear program for static appointment scheduling capturing no-show behavior; and subsequently, a dynamic scheduling using a multi-stage stochastic linear program. In research by [30], the stochastic overbooking in outpatient appointment scheduling for clinical use was modeled. Patient wait time, provider over time and patient revenue were considered in the objective function of the model formulation. In this research, we devise a two-stage stochastic mixed-integer program for appointment scheduling while considering uncertainties in patient’s demand and no-show behavior of patients. The model also considers fair and equitable and non-appointment slot assignments for each provider. We consider an OB/GYN clinic for our case-study, and different types of services and categories are as shown in Table 1.

### Table 1. OB/GYN patient types: OB - Obstetrics; GYN - Gynecology, MAU - Monthly active users.

| Service Category | Service Type                  |
|------------------|-------------------------------|
| Low Risk OB      | New Low-Risk OB               |
|                  | Follow-Up Low-Risk OB         |
| High Risk OB     | New GYN                       |
| Gynecology       | MAU GYN                       |
|                  | Established GYN               |
|                  | Results GYN                   |

### III. NOTATION AND MODEL FORMULATION

We design an appointment scheduling model capturing multi-type patient channeling to different provider levels in an OB/GYN clinical specialty. Based on the research on OB/GYN specialty clinics by [33] and [35], we divide patients into three categories, and consequently, seven patient types.

Let $N$, $R$ and $T$ be the sets (indexed by $n$, $r$, $t$) denoting patient categories, providers and planning horizon, respectively. The set $H$ represents the total available slots for a provider within the planning horizon, and $N'$ is the set of new patient categories which should have fair and equitable assignments to the providers. Each day in the planning horizon is divided into 16 fifteen-minutes slots. Service time duration for each patient type is based on the literature on OB/GYN clinic ([33], [35]). The authors in [35] collected data from the West Little Rock (WLR) clinic operated under the University of Arkansas for Medical Sciences (UAM). The parameters $a_r$ and $p_r$ denote the target for new patient category and penalty for the deviation from the target for a provider $r$. The average service time for a new patient is defined as penalty. Parameters $c_i$ and $F$ represent the criticality for a patient category $i$ and overall criticality factor allowed in a slot, respectively, and $M$ is a large constant. The parameter $v_r$ represents the total number of available slots for a provider $r$ within the entire planning horizon $H$.

Let $\omega$ be a multi-variate random variable representing the uncertainties in call-volume and no-show behaviors of the patients.

Two-stage stochastic programming is a common approach for modeling decision making problems that involve uncertainty. First-stage decision variables represent ‘here-and-now’ decisions that are determined before the realization of randomness, and second-stage decisions are determined after scenarios representing uncertainties are presented. In this two-stage formulation, the first-stage decides the scheduling template for the providers, which indicates the number of patient categories to be assigned for each slot. The first-stage scheduling template is exposed to multiple scenarios of patient call volumes in the second-stage where the indirect wait time is minimized. We minimize the deviation of new-patient type of each provider from the given target in the first-stage, and expected indirect wait time in the second-stage. In the first-stage, the integer variable $x_{irt}$ denotes the number of patient category $i$ assigned to provider $r$ per time-slot $t$, variable $e_r$ measures the deviation of the assignments for a provider $r$ from its target, and $z_{rt}$ is a binary variable indicating whether a slot has been used by the provider $r$ or not.

In the second-stage, let $P$ be the set of patients, indexed by $p$, and $p(i)$ and $p(d)$ represent patient’s category and desired slot of appointment, respectively. The parameter $w_i$ represents a penalty weight for the patient category $i \in N$, and $\epsilon$ is a parameter in the range $(0,1)$. The no-show parameter for a patient $p(i)$ for a scenario $\omega$ is represented as a binary vector $\psi(i)(\omega)$, where a value of one represents that the patient will not show up for the appointment, and zero for otherwise. The second-stage binary variable $y_{p(i)r}(\omega)$ represents the assignment of patient $p(i)$ with a provider $r$ for the time slot $t$ in a scenario $\omega$. The first-stage objective function minimizes the deviation from the proportion of new patients allotted to each provider, and second-stage objective function minimizes the expected indirect wait times for the patients.

The first-stage formulation is represented as follows:

Min $\sum_{r \in R} p_r e_r + \mathbb{E}[\psi(x, \omega)]$ \hspace{1cm} (1)

s.t. $e_r + \sum_{i \in T} \sum_{n \in N'} x_{irt} \geq a_r \hspace{0.5cm} \forall r \in R$, \hspace{1cm} (2)

$e_r - \sum_{r \in R} \sum_{n \in N'} x_{irt} \geq -a_r \hspace{0.5cm} \forall r \in R$, \hspace{1cm} (3)
\[ \sum_{i \in N} c_i x_{irt} \leq F \quad \forall r \in R, t \in T, \]  
\[ \sum_{i \in N} x_{irt} \leq M (1 - z_{rt}) \quad \forall r \in R, t \in T, \]  
\[ \sum_{i \in N} z_{rt} \leq v_r \quad \forall r \in R, H \subset T, \]  
\[ x_{irt} \in \mathbb{Z}^+, e_r \geq 0, z_{rt} \in \{0, 1\}. \]

For a given first-stage solution \( x \) and a realization \( \omega \) of the random variable \( \tilde{\omega} \), the second-stage recourse function is defined as follows:

\[ \text{Min} \quad \varphi(x, w) = \sum_{p(i) \in P} \sum_{r \in R} w_i(t - p(d))^{(1+\epsilon)} y_{p(i)rt}(\omega) \]

\[ \text{s.t.} \quad \sum_{p(i) \in P} (1 - s_{p(i)}(\omega)) y_{p(i)rt}(\omega) \leq x_{irt} \quad \forall i \in N, r \in R, t \in T, \]  
\[ \sum_{i \in T} (\sum_{p(d) \leq t} r \in R) = 1 \quad \forall i \in N, p(i) \in P : s_{p(i)} = 0, \]  
\[ y_{p(i)rt}(\omega) \in \{0, 1\} \quad \forall i \in N, p(i) \in P, r \in R, t \in T. \]

In the above formulation, constraints (2) and (3) capture the deviation for total number of appointments for the new patients for each provider \( r \). The lower or upper deviation from the target \( a_r \) is captured using the variable \( e_r \) and is penalized in the objective function (1). Constraint (4) is a capacity restriction which limits the number of different patient categories assigned in a slot, and constraint (5) defines the indicating relationship between the variables \( x \) and \( z \). Considering other non-clinic activities, the total availability for a provider is defined in the constraint (6). Finally, restrictions for the variables are defined in the constraints (7).

In the second-stage formulation, constraints (9) state that individual assignment of patients depends upon the scheduling template generated in the first-stage. Constraints (10) assure that a patient gets an appointment within the planning horizon. For the objective, a super-linear function \((t - p(d))^{(1+\epsilon)}\) is used so that indirect delays among the patients are fair and equitable, and the function is weighted by the parameter \( w_i \) based on the patient category \( i \). The second-stage objective function (8) represents a weighted indirect wait time for the patients for a given scenario \( \omega \).

**A. CLINIC PATIENT FLOW SIMULATION**

The patient flow simulation for the clinic is executed for each day based on the scheduling template proposed by the optimization model described in the previous section. Patients’ direct and indirect wait times are measured. The clinic patient flow involves two stages: time with the nurse and time with the provider. Once a patient walks into the clinic, the patient waits in the lobby for the nurse to become available. After being visited by the nurse, the patient waits for the provider in the exam room. We assume that patients who are scheduled for a particular day must be served before the end of that day even if the provider has to work overtime. Patients are assumed not to leave the exam room until the provider finishes all required tasks. Also, patients may arrive late for their appointments. By arriving late, patients may increase the wait times for the patients that follow them. Therefore, we assume that patients are called in the order of their arrival time.

The first-stage solution \( x \) is evaluated in a simulation model for a clinic, and the average direct wait time for the patients in each slot is noted. Whenever the direct wait time for a slot is beyond a given user threshold of twenty minutes, the assignment sequence for the \( x \) variables is prohibited by adding the constraints (12)-(15) for the next iteration of branch and cut procedure. Let \( L \) be the set of integers \( \{1, 2, 3, 4\} \), and a parameter \( m_l \in H \). Similarly, \( G \) is the set of consecutive time-periods where the direct wait time from the simulation is beyond the user’s threshold. The binary variable \( \bar{x}_{irt} \) is used to discretize the variable \( x \). The constraints added for the direct wait time are as follows:

\[ x_{irt} = \sum_{l \in L} m_l \bar{x}_{irt} \quad \forall i \in N, r \in R, t \in T, \]  
\[ \sum_{l \in L} \bar{x}_{irt} \leq l \quad \forall i \in N, r \in R, t \in T, \]  
\[ \sum_{i \in N, l \in L, t \in G} \bar{x}_{irt} \leq |G| - 1, \quad \forall r \in R, \]  
\[ \bar{x}_{irt} \in \{0, 1\}. \]

Constraints (12) - (13) discretize the variable \( x \) to identify the exact combination of number of patients in different types that are scheduled for a set of consecutive time slots \( G \). Constraints (14) denote that such combination should be avoided in the next iteration of branch and cut algorithm. It should be noted that the constraints (12)-(15) are dynamically added, i.e., every feasible \( x \) is evaluated in a clinic simulation model during the branch and cut procedure, and whenever the threshold for direct wait time is violated, the respective constraints are added. For every combination (denoted by set \( G \)) that exceeds the given threshold, the corresponding set of constraints (12) - (14) are generated and added to the model for the next iteration of branch and cut procedure.

Algorithm 1 presents pseudocode for our approach based on the two-stage optimization model and simulation. The two-stage stochastic programming model without direct wait time is solved, and a scheduling template is obtained from the first-stage solution. Using the scheduling template, each day in the planning horizon is simulated using given input with 200 replications, and specific percentiles of patient flow measures are calculated. If any of the measures violate the patients’ direct wait time thresholds, a new set of constraints (12), (13), and (14) are added to the model, which is then re-optimized. The process is repeated until the optimization
Algorithm 1: Clinic Patient Flow Simulation

Input: Output of the two-stage stochastic model (2)-(11)
Output: Constraints (12)-(14)

Initialization:
1. Optimize the two-stage stochastic model (2)-(11);
2. for $t$ in $T$ do
3. run simulation model for each day in the planning horizon;
4. if (estimates ≥ any of the thresholds) then
5. construct the corresponding constraints (12)-(14);
6. end if
7. end for
8. return constraints (12)-(14)

The call center simulation uses either “first-call-first-service scheduling policy” (FCFS) or the scheduling template’s allocation from the two-stage stochastic programming model described in the earlier section. A FCFS is used to allot or cancel patients’ appointments. The scheduler at the call center estimates the priority of available appointment slots that should be offered to a patient based on her desired date and patient type. When a patient requests an appointment for a desired date and slot, the policy allocates a slot based on its feasibility and proximity to the desired date. For feasibility check, the current total criticality factor is calculated with the new assignment, and it should be less than the allowed threshold. Similarly, when the scheduling template is used, an assignment is made based on the capacity allocated for each patient type in the template. Feasibility check is not required for scheduling template as it is accounted in the two-stage stochastic programming model via constraint (4). To generate a patient’s choice regarding accepting an appointment, a random number from the uniform distribution $U(0, 1)$ is generated and compared to an “acceptance” threshold. If the random number is less than this threshold, the patient accepts the corresponding appointment slot; otherwise, another slot is offered and the process continues until the patient accepts. Appointment cancellations are handled similarly: a random number is generated from $U(0, 1)$, and if the random number is less than the clinic’s cancellation rate, the patient or clinic cancels the appointment and the patient is removed from the scheduling grid.

B. CALL CENTER SIMULATION

To evaluate the efficiency of the two-stage scheduling template, we simulate the practice at the appointment call center. The call center simulation uses either “first-call-first-service scheduling policy” (FCFS) or the scheduling template’s allocation from the two-stage stochastic programming model described in the earlier section. A FCFS is used to allot or cancel patients’ appointments. The scheduler at the call center estimates the priority of available appointment slots that should be offered to a patient based on her desired date and patient type. When a patient requests an appointment for a desired date and slot, the policy allocates a slot based on its feasibility and proximity to the desired date. For feasibility check, the current total criticality factor is calculated with the new assignment, and it should be less than the allowed threshold. Similarly, when the scheduling template is used, an assignment is made based on the capacity allocated for each patient type in the template. Feasibility check is not required for scheduling template as it is accounted in the two-stage stochastic programming model via constraint (4). To generate a patient’s choice regarding accepting an appointment, a random number from the uniform distribution $U(0, 1)$ is generated and compared to an “acceptance” threshold. If the random number is less than this threshold, the patient accepts the corresponding appointment slot; otherwise, another slot is offered and the process continues until the patient accepts. Appointment cancellations are handled similarly: a random number is generated from $U(0, 1)$, and if the random number is less than the clinic’s cancellation rate, the patient or clinic cancels the appointment and the patient is removed from the scheduling grid.

IV. CASE STUDY

In this section, we present a case study to demonstrate the performance of the proposed two-stage stochastic programming and simulation models. We consider a multi-category outpatient appointment scheduling for a women’s OBGYN clinic. The characteristics and demand for different patient types for the clinic used in the case study are acquired from the literature of women’s OBGYN specialty clinics.

A. DATA AND STUDY DESIGN

Based on the literature on OB/GYN clinics, the common issue is related to time, equipment, and exam rooms scheduled for several service categories. Each patient type needs specific services such as prenatal and follow up care of pregnancy and miscarriage. The study in [33] has divided patient types with respect to the needed service type into three categories (Table 1). In this case study, each clinic session is defined as a day and is divided into 16 time-slots with identical service time duration of 15 minutes. There are two providers available on all days of the week who can provide all service categories for different patient types. Patients are scheduled with any available provider in each clinic session (morning/afternoon). Service time duration for different servers in the clinic such as time spent with nurse and provider are based on [33] and [35].

We assume there are two sessions: morning and afternoon, and each session has eight time-slots. OB/GYN clinics services rendered for different patient types are considered in different sessions of a week based on the study in [33]. Also, typically women clinics consider appointment scheduling for all providers in a clinic and not for specific ones. Therefore, patients can be seen by any available provider upon their appointment time. Since multiple providers are assigned for each day, overbooking is allowed for each time slot. The parameters used for the study are shown in Table 2. Uniform distribution was used to obtain the service time for different patient categories. The seven patient types are organized under three categories, their respective distributions for the service times are provided in the table. Similarly, a different no-show rate was used different patient types and this was used to populate the no-show rate parameter $s_{p(i)}(\omega)$ in the optimization model. The total number of patients were divided into three categories of patients in the proportion of 30%, 30%, and 40% for Low Risk OB, High Risk OB, and Gynecology, respectively. The segregation of categories is used to determine the number of patient types within each category. Similarly, the criticality factors for the patient types are 1.669, 0.39, and 0.67 for the first provider, and 1.2, 0.869, and 0.669 for the second provider. The total criticality factor for any time slot used for the computational experiments is 1.99. Based on the capacity, the total number of calls is 350 for the study. Also, for the study ‘Low Risk OB’ patient category was categorized as new patient type. The desired day for an appointment by the incoming calls of the patients were in proportion of 20% each for Monday, Tuesday and Thursday, 30% for Wednesday, and finally, 10% for Friday. The assumption is that the most of the appointment requests fall during the middle of the week and it is substantially lower for the last day of the week. For the distribution of the calls, a decision is made for the day of the call based on
the proportion. Then a uniform distribution \( U(1, 16) \) is used to assign it to a slot within a day. We used six scenarios for the second-stage, and a standard deviation of three slots \( U(-3, 3) \) around the average is used as distribution for creating the scenarios. The task of the first-stage model is to determine the scheduling template for the providers while minimizing the deviation and expected costs from first- and second-stages, respectively.

**B. OPTIMAL WEEKLY SCHEDULING TEMPLATE**

The output of two-stage stochastic model presented in the earlier section is a scheduling template (SC-T). The scheduling template is used as a guideline in the call center to assign appointments for incoming calls from the patients. The assignment pattern using scheduling template is compared with the ‘FCFS’ policy where the assignments are made solely based on the capacity of a desired slot of the patient. The SC-T is derived based on the costs of second-stage scenarios where the potential call pattern is taken into account. However, FCFS policy is short sighted where it looks only the immediate requirement. Sample weekly scheduling templates are represented in the Tables 3 and 4. In the tables, NL, FL, FH, NG, MG, EG, and RG represent New Low-Risk OB, Follow Up Low-Risk OB, Follow Up High-Risk OB, New GYN, MAU GYN, Established GYN, and Results GYN, respectively. When the direct wait time is lenient, there are more patients scheduled for a given time slot. This is depicted in the tables, where Table 4 is more lenient in direct wait time for the patients in clinic compared to the threshold used for Table 3. During the call center simulation, appointments for the incoming calls from the patients are made based on the number and type of patient types assigned in a slot. If a particular type is not available then the nearest unused slot with the requested patient type is assigned. However with the ‘FCFS’ policy, the assignments are solely based on the capacity of a slot, i.e., whether by adding another patient will exceed the total criticality factor or not.

**C. COMPUTATIONAL RESULTS AND INSIGHTS**

The computational experiments were performed on a Dell computer with Xeon 2.60 GHz and 80 GB RAM. The algorithm was implemented using Java programming language, and CPLEX 12.10 is used as a solver. The two-stage model presented in the constraints (1)-(11) along with the dynamic direct wait time constraints (12)-(15) is referred as ‘SMILP’, and the scheduling template obtained by solving the model is referred as ‘SC-T’. The deterministic equivalent problem (DEP) for the two-stage formulation was solved using CPLEX 12.10 as optimization solver. Dynamic direct wait time constraints are added to the first-stage model via ‘LazyConstraintCallback’ feature of CPLEX. Whenever there is a first-stage feasible integer solution \( \bar{x} \) for SMILP, the callback function runs the clinic simulation with \( \bar{x} \). A threshold of 20 minutes is used for the study. Whenever the direct wait time in a slot is beyond 20 minutes in clinic simulation, the constraints (12)-(15) are generated for the slot and the one which is adjoining it, and added to the model. This way, the combination in \( \bar{x} \) for the slot will be avoided in the next iteration. In a conventional comparison, the objective function values from the model and ‘FCFS’ policy are calculated and a value of stochastic solution is estimated. Instead of this, we simulated the clinic and call center operations, and used the policies to compare the utilization of providers, direct and indirect wait times. This provides a more realistic way to measure the performance of the policies. We conducted eight simulation runs, and each run was repeated five times. 

Figures 2 and 3 compare the performance of direct and indirect wait times between two-stage model and FCFS.
TABLE 4. Sample of a scheduling template-II.

| Slot Index | Mon       | Tue       | Wed       | Thu       | Fri       |
|------------|-----------|-----------|-----------|-----------|-----------|
| 1          | FL-FL-FL  | FH-FH-EG  | FH-FH-PH  | FH-NL     | PH-FH     |
| 2          | FL-FL-FH  | FL-FL-MG  | FL-FH-EG  | FH-NL     | FH-FH-FL  |
| 3          | FL-FI-FH  | FL-FH-FL  | FL-FH-FL  | FL-NL     | EG-FI     |
| 4          | FL-FI-EG  | FH-FE-EG  | MG-FE-EG  | FH-FI     | FH-FI-FL  |
| 5          | FH-FI-MG  | FH-NL-MG  | FH-FI-EG  | FH-NL     | FH-FI-NG  |
| 6          | FH-FI-PH  | FH-FI-EG  | FH-FH-NL  | NL-EG     | FH-FI-NG  |
| 7          | FH-EG-EG  | FH-FH-EG  | FH-FH-NL  | FH-FI-EG  | FH-FI-NL  |
| 8          | FH-EG-EG  | FL-FH-FL  | FH-FH-EG  | FH-NL     | FH-FI-NG  |
| 9          | FL-FI-MG  | FL-FI-EG  | FL-FI-EG  | FH-NL     | FH-FI-FL  |
| 10         | FL-EG     | NL-FI-FL  | FH-NG-EG  | NL-FE     | FH-EG-FL  |
| 11         | FL-FL-FI  | FL-FI-FL  | FL-FI-EG  | FL-NL     | FH-FI-FL  |
| 12         | FH-MG     | FH-EG-MG  | FH-FI-EG  | FH-FI     | FH-FI-NG  |
| 13         | FH-FI-MG  | FL-FI-FL  | FL-FI-NG  | FH-FI     | FH-FI-NG  |
| 14         | FL-FL-FI  | FL-FI-FL  | FL-FI-NG  | FH-FI     | FH-FI-NG  |
| 15         | FL-FI-FH  | FL-FI-FL  | FL-FI-NG  | FL-FI-MG  | FL-FI-NG  |
| 16         | FH-FI-NG  | NL-FI-FL  | FL-FI-EG  | FL-FI-NG  | FREE      |

FIGURE 2. Comparison of indirect wait time between two-stage model and index policy.

FIGURE 3. Comparison of direct wait time between two-stage model and index policy.

FIGURE 4. Direct wait time constraints for the two-stage model using simulation model.

In terms of distribution of indirect wait time for the patients in a replication, the averages were three and ten slots for SC-T and FCFS policy, respectively. Especially, it should be noted that only few patients had non-zero indirect wait time using the scheduling template compared to more than 70% using FCFS policy. Table 5 summarizes the results from the simulation study. The reported average values are for each patient. AIWT C-1 refers to average indirect wait time for patient category-1, similarly AIWT C-2 and C-3 refer to average indirect wait time for patient category-2 and category-3. The average direct wait time had the biggest benefit by using SC-T policy, respectively. In the figures, the horizontal axis denotes the indices for the run where each run had five replications, and the vertical axis denotes the wait time in slots. On an average, each patient had about three slots of indirect wait time from SMILP compared to ten units for using FCFS. Similarly, for the direct wait time, the SMILP model had insignificant wait time (less than a slot) compared to an average of 1.5 slots for the policy. This significant improvement is due to the dynamic addition of constraints (12)-(15) in the SMILP model. The number of such constraints added for each run is shown in Figure 4. In the figure, ‘DW Cuts’ refer to number of direct wait constraints. On an average, there were 115 such constraints added dynamically to the first-stage of SMILP model for each run.
as the constraints were generated by the simulation model and added directly to the SMILP model.

V. CONCLUSION AND FUTURE DIRECTIONS

Outpatient access to the facilities is controlled through appointment scheduling. Patients often suffer from significant delays in obtaining appointments and wait time within the clinic. A well-established appointment scheduling system can help clinics reduce patients’ indirect wait time while also improving patient flow within the clinic. Clinic managers have to handle multiple issues when scheduling patient appointments. While different patient types have different complexities, there are uncertainties in the pattern of calls for appointments and patients’ willingness to wait for an appointment. Some patients may call in advance to book appointments, while others ask for same-day appointments. This study proposes solving the appointment scheduling problem with a two-stage stochastic programming model integrated with a simulation model to minimize patients’ indirect wait time for an appointment while maintaining a patient flow in the clinic that is within the constraints of acceptable patient and practitioner performance thresholds. The model proposes a patient scheduling template for the call center in order to help the clinic manager to reduce the patient appointment delays, scheduling errors and improve the efficiency of resource allocation.

We use a numerical case study inspired by a real-world healthcare system to validate our proposed approach instead of a first-call-first-serve approach. Our model also performs better when the provider is assigned greater capacity for other clinical activities. Determining the optimal schedule for re-running the optimization model based on changes in the uncertainties is an important avenue for future research. Moreover, while our two-stage stochastic programming model determines the number of each patient type that can be scheduled in each appointment slot, it does not provide any guidance regarding when the call center should offer each open appointment to a patient. Integrating the call center scheduling process with our two-stage stochastic programming approach could result in further improving patients’ indirect wait times. In the proposed risk-neutral two-stage stochastic model, we consider expected value in the second-stage function. For a future extension, a mean-risk performance measure like Conditional-Value-at-Risk (CVaR) can be utilized to estimate the robustness of the scheduling template.

### TABLE 5. Characteristics of the results.

|               | SC-T Policy |
|---------------|-------------|
| Average indirect wait time (AIWT) (slots) | 3.17 10.28 |
| AIWT C-1 (slots) | 0.84 5.94   |
| AIWT C-2 (slots) | 0.94 1.24   |
| AIWT C-3 (slots) | 1.39 3.10   |
| Average direct wait time (slots) | 0.04 1.54   |
| Average direct wait time cuts | 115 -       |

### REFERENCES

[1] S. M. Asch, E. A. Kerr, J. Keevey, J. L. Adams, C. M. Setodji, S. Malik, and E. A. McGlynn, “Who is at greatest risk for receiving poor-quality health care?” New England J. Med., vol. 354, no. 11, pp. 1147–1156, Mar. 2006.
[2] M. S. Donaldson, J. M. Corrigan, and L. T. Kohn, To Err is Human: Building a Safer Health System, vol. 6. Washington, DC, USA: National Academies Press, 2000.
[3] A. Baker, Crossing the Quality Chasm: A New Health System for the 21st Century, vol. 323, no. 7322. London, U.K.: British Medical Journal Publishing, Institute of Medicine, 2001, p. 1192.
[4] M. A. Schuster, E. A. McGlynn, and R. H. Brook, “How good is the quality of health care in the united states?” Milbank Quart., vol. 76, no. 4, pp. 517–563, Dec. 1998.
[5] N. S. Wenger, D. H. Solomon, C. P. Roth, C. H. MacLean, D. Saliba, C. J. Kanber, L. Z. Rubenstein, R. T. Young, E. M. Sloss, R. Louie, and J. Adams, “The quality of medical care provided to vulnerable community-dwelling older patients.” Ann. Intern. Med., vol. 139, no. 9, pp. 740–747, 2003.
[6] R. M. Poses, “A cautionary tale: The dysfunction of American health care,” Eur. J. Intern. Med., vol. 14, no. 2, pp. 123–130, Mar. 2003.
[7] Centers for Medicare & Medicaid Services. Health Insurance Coverage in the United States: 2018. Accessed: Apr. 5, 2020. [Online]. Available: https://www.census.gov/library/publications/2019/demo/p60-267.html
[8] E. R. Berchick, J. C. Barnett, and R. D. Upton. Health Insurance Coverage in the United States: 2018. Accessed: Apr. 5, 2020. [Online]. Available: https://www.census.gov/content/dam/Census/library/publications/2019/demo/p60-267.pdf
[9] S. S. Teleki, M. E. Sorbero, L. Hibborne, S. Lovejoy, and L. Bradley. Performance Measurement in the Hospital Outpatient Setting. [Online]. Available: https://aspe.hhs.gov/report/performance-measurement-hospital-outpatient-setting/.
[10] MedPAC, “Section 1, national health care spending and medicare spending,” in A Data Book: Healthcare Spending and the Medicare Program. Washington, DC, USA: MedPAC, 2007.
[11] MedPAC, “Section 8, ambulatory care: Physicians, hospital outpatient services, ambulatory surgical centers and imaging services,” in A Data Book: Healthcare Spending and the Medicare Program. Washington, DC, USA: MedPAC, 2007.
[12] P. Santibáñez, V. S. Chow, J. French, M. L. Puterman, and S. Tyldesley, “Reducing patient wait times and improving resource utilization at British Columbia cancer agency’s ambulatory care unit through simulation,” Health care Manage. Sci., vol. 12, no. 4, p. 392, 2009.
[13] N. T. J. Bailey, “A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times,” J. Roy. Stat. Soc. B, Methodol., vol. 14, no. 2, pp. 185–199, Jul. 1952.
[14] T. Cayirlı, E. Veral, and H. Rosen, “Designing appointment scheduling systems for ambulatory care services,” Health Care Manage. Sci., vol. 9, no. 1, pp. 47–58, Feb. 2006.
[15] K. J. Klassen and R. Yoogalingam, “Improving performance in outpatient appointment services with a simulation optimization approach,” Prod. Oper. Manage., vol. 18, no. 4, pp. 447–458, Jul. 2009.
[16] G. C. Kaandorp and G. Koole, “Optimal outpatient appointment scheduling,” Health Care Manage. Sci., vol. 10, no. 3, pp. 217–229, Jun. 2007.
[17] R. Alizadeh, J. Rezaeian, M. Abedi, and R. Chiong, “A modified genetic algorithm for non-emergency outpatient appointment scheduling with highly demanded medical services considering patient priorities,” Comput. Ind. Eng., vol. 139, Jan. 2020, Art. no. 106106.
[18] L. Leaven and X. Qu, “Improving appointment scheduling systems in outpatient clinics using a deterministic modeling approach (MILP),” in Proc. IEEE Annu. Conf. Reno, NV, USA: ProQuest, 2011.
[19] W. M. Macharia, “An overview of interventions to improve compliance with appointment keeping for medical services,” J. Amer. Med. Assoc., vol. 267, no. 13, pp. 1813–1817, Apr. 1992.
[20] B. P. Berg, B. T. Denton, S. Ayca Erdogan, T. Rohleder, and T. Huschka, “Optimal booking and scheduling in outpatient procedure centers,” Comput. Oper. Res., vol. 50, pp. 24–37, Oct. 2014.
[21] J. A. Deife, C. Z. Conklin, J. M. Smith, and J. Poole, “Psychotherapy appointment no-shows: Rates and reasons,” Psychotherapy, Theory, Res., Pract., Training, vol. 47, no. 3, p. 413, 2010.
[22] J. Feldman, N. Liu, H. Topaloglu, and S. Ziya, “Appointment scheduling under patient preference and no-show behavior,” Oper. Res., vol. 62, no. 4, pp. 794–811, Aug. 2014.
[23] J. Rau, “Medicare to begin basing hospital payments on patient-satisfaction scores,” *Kaiser Health News*, vol. 28, Apr. 2011. [Online]. Available: https://khn.org/news/medicare-hospital-patient-satisfaction/

[24] D. S. Festinger, R. J. Lamb, D. B. Marlowe, and K. C. Kirby, “From telephone to office: Intake attendance as a function of appointment delay,” *Addictive Behav.*, vol. 27, no. 1, pp. 131–137, 2002.

[25] J. Dreher, M. Froimovici, Y. Bibi, D. A. Vardy, A. Cicurel, and A. D. Cohen, “Nonattendance in obstetrics and gynecology patients,” *Gynecol. Obstetric Invest.*, vol. 66, no. 1, pp. 40–43, 2008.

[26] Merritt Hawkins Associates. (Sep. 2017). *Survey of Physician Appointment Wait Times*. [Online]. Available: https://www.merritthawkins.com/uploadedFiles/MerrittHawkins/Content/Pdf/mha2017waittimesurveyPDF.pdf

[27] D. Gupta and B. Denton, “Appointment scheduling in health care: Challenges and opportunities,” *IIE Trans.*, vol. 40, no. 9, pp. 800–819, Jul. 2008.

[28] T. Cayirli and E. Veral, “Outpatient scheduling in health care: A review of literature,” *Prod. Oper. Manage.*, vol. 12, no. 4, pp. 519–549, Jan. 2009.

[29] N. Liu, S. Ziya, and V. G. Kulkarni, “Dynamic scheduling of outpatient appointments under patient no-shows and cancellations,” *Manuf. Service Oper. Manage.*, vol. 12, no. 2, pp. 347–364, Apr. 2010.

[30] K. Muthuraman and M. Lawley, “A stochastic overbooking model for outpatient clinical scheduling with no-shows,” *IIE Trans.*, vol. 40, no. 9, pp. 820–837, Jul. 2008.

[31] S. A. Erdogan, A. Gose, and B. T. Denton, “Online appointment sequencing and scheduling,” *IIE Trans.*, vol. 47, no. 11, pp. 1267–1286, Nov. 2015.

[32] S. A. Erdogan and B. Denton, “Dynamic appointment scheduling of a stochastic server with uncertain demand,” *Informs J. Comput.*, vol. 25, no. 1, pp. 116–132, Feb. 2013.

[33] X. Qu, Y. Peng, N. Kong, and J. Shi, “A two-phase approach to scheduling multi-category outpatient appointments—A case study of a women’s clinic,” *Health Care Manage. Sci.*, vol. 16, no. 3, pp. 197–216, Sep. 2013.

[34] C. Zacharias and M. Armony, “Joint panel sizing and appointment scheduling in outpatient care,” *Manage. Sci.*, vol. 63, no. 11, pp. 3978–3997, Nov. 2017.

[35] R. B. Lenin, C. L. Lowery, W. C. Hitt, N. A. Manning, P. Lowery, and H. Eswaran, “Optimizing appointment template and number of staff of an OB/GYN clinic—micro and macro simulation analyses,” *BMC Health Services Res.*, vol. 15, no. 1, p. 387, Jun. 2015.

**SAMIRA FAZEL ANVARYAZDI** received the M.S. degrees in pure mathematics and mathematical statistics and the Ph.D. degree in industrial and systems engineering from Wayne State University. She is currently a Lecturer of data analytics with the Olin Business School and a Faculty Scholar with the Institute for Public Health, Washington University in Saint Louis. Her research interests include stochastic programming, robust optimization techniques, data science, and compartmental models with applications in healthcare systems engineering and business problems.

**SARAVANAN VENKATACHALAM** received the M.S. and Ph.D. degrees in industrial and systems engineering from Texas A&M University, College Station, in 2003 and 2014, respectively. He is currently an Assistant Professor of industrial and systems engineering with Wayne State University. His research interests include stochastic programming, large-scale optimization, and discrete event modeling and simulation, and applications of interest include supply chain management, healthcare, pricing and revenue management, and energy management.

**RATNA BABU CHINNAM** received the M.S. and Ph.D. degrees in industrial engineering from Texas Tech University, Lubbock, in 1990 and 1994, respectively. He is currently a Professor of industrial and systems engineering and the Director of the Big Data and Business Analytics Group, Wayne State University. His research interests include business analytics, big data, supply chain management, freight logistics, operations management, sustainability, healthcare systems engineering, and smart engineering systems.

---

**SAMIRA FAZEL ANVARYAZDI** received the M.S. degrees in pure mathematics and mathematical statistics and the Ph.D. degree in industrial and systems engineering from Wayne State University. She is currently a Lecturer of data analytics with the Olin Business School and a Faculty Scholar with the Institute for Public Health, Washington University in Saint Louis. Her research interests include stochastic programming, robust optimization techniques, data science, and compartmental models with applications in healthcare systems engineering and business problems.

**SARAVANAN VENKATACHALAM** received the M.S. and Ph.D. degrees in industrial and systems engineering from Texas A&M University, College Station, in 2003 and 2014, respectively. He is currently an Assistant Professor of industrial and systems engineering with Wayne State University. His research interests include stochastic programming, large-scale optimization, and discrete event modeling and simulation, and applications of interest include supply chain management, healthcare, pricing and revenue management, and energy management.

**RATNA BABU CHINNAM** received the M.S. and Ph.D. degrees in industrial engineering from Texas Tech University, Lubbock, in 1990 and 1994, respectively. He is currently a Professor of industrial and systems engineering and the Director of the Big Data and Business Analytics Group, Wayne State University. His research interests include business analytics, big data, supply chain management, freight logistics, operations management, sustainability, healthcare systems engineering, and smart engineering systems.