1-1-2018

Managing Operational Efficiency And Health Outcomes At Outpatient Clinics Through Effective Scheduling

Samira Fazel Anvaryazdi
Wayne State University,

MANAGING OPERATIONAL EFFICIENCY AND HEALTH OUTCOMES AT OUTPATIENT CLINICS THROUGH EFFECTIVE SCHEDULING

by

SAMIRA FAZEL ANVARYAZDI

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2018

MAJOR: INDUSTRIAL ENGINEERING

Approved by:

_____

Primary-major Advisor Date

_____

Co-major Advisor Date

_____

_____

_____

DEDICATION

To my parents and family, for all their love and support

ACKNOWLEDGMENTS

I would like to thank, first and foremost, my advisors, Dr. Ratna Babu Chinnam and Dr. Saravanan Venkatachalam for all their continuous support and guidance throughout my Ph.D. Their inspiring and thought-provoking directions have significantly improved my academic self. Their insightful comments about my dissertation helped me a lot to shape my research. It would not have been possible to complete this Ph.D. without them. I would also like to extend a thank you to my dissertation committee: Dr. Leslie Monplaisir, Dr. Qingyu Yang and Dr. Robert Reynolds for great recommendations and valuable support during my dissertation. Last, but not least, I would like to thank my family, in particular my parents, my hero and angel, who have given me their unconditional love and support. Without them, I would have quit long ago.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

The goal of health care is to maintain and improve health through the prevention, diagnosis, and treatment of physical and mental disease in human beings. A health care system is the organization of resources that deliver health care services to the target populations who are in need of health care. In particular, delivery of the health care is supported by health professionals, such as providers, in related health categories such as medicine, psychology, physical therapy, OBGYN clinics and other health professions which are all part of health care. It includes work done in providing primary care, secondary care, and tertiary care, as well as in public health.

A variety of studies have documented the substantial deficiencies in the quality of health care delivered across the United States(Asch et al. 2006); (Kohn et al. 2000); (Anon 2001); (Schuster et al. 1998); (Wenger et al. 2003). Attempts to reform the United States health care system in the 1980s and 1990s were inspired 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, access, quality, and costs are still challenges, and medical professionals are increasingly dissatisfied (Poses 2003). According to the Centers for Medicare and Medicaid Services (CMS 2016), costs associated with national health care increased 4.3 percent in 2016 compared to 5.8 percent growth in 2015. United States devotes 17.9% of GDP to health care (spending $10,348 per person, in 2016, or $3.3 trillion total), compared with 9% in Britain, yet life expectancy is slightly below average for a rich country and nearly 50 million Americans were uninsured in 2012 (CMS 2016). While there are no comparable studies for the quality of care delivered in the hospital outpatient setting, pervasive deficits across the health system suggest existence of the similar problems, particularly since 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, (Teleki et al. 2007). In addition to the potential quality of care deficits in the hospital outpatient setting, the Centers for Medicare and Medicaid Services (CMS) and others have also

observed growth in the volume of services and costs for care delivered in this setting. Outpatient clinics such as Diabetes, OBGYN, and cancer treatment centers represent a unique, but growing, point of care in the United States health care delivery system.

In recent years, appointment scheduling in outpatient clinics has attracted much attention in health care delivery systems. Increase in demand for health care services as well as health care costs are the most important reasons and motivations for health care decision makers to improve health care systems. The goals of health care systems include patient satisfaction as well as system utilization. According to (Gupta & Denton 2008), less attention goes to the benefit of patients compared to that of clinic services and providers. As a result, health care systems have recently set goals regarding patient satisfaction and improving the performance of the health system by timely and appropriate health care delivery. (Liu et al. 2010) and (Gupta & Denton 2008) have reported that parameters such as demand uncertainty, patient no-show behavior, patient/provider unpunctuality, stochastic servers and multiple patient types such as real situations, modeling approaches, and solution methodologies are the criteria most commonly used in appointment scheduling, which makes it challenging. Many studies have documented the no-show rate in medical practice. (Macharia et al. 1992) reported a 42% average no-show rate which ranges from 6% to 92% in outpatient clinics. (Berg et al. 2014) reported 13% to 24% no-show rates at endoscopy clinics for different service types. (Festinger et al. 2002) shows post intervention no-show rates ranging from 28% to 45%. (Dreiher et al. 2008) results show the overall no-show rate at OBGYN clinics as 30.1%. They investigated the strong relationship between patients' appointment delays and no-show cases in OBGYN specialty clinics. In psychotherapy appointments, a 21% no-show rate was reported by (DeFife et al. 2010).

(Ahmadi-Javid et al. 2017) suggest that decision making in outpatient appointment scheduling can be classified into three categories: strategic, tactical, and operational decisions which are long, medium and short-term decisions, in that order. The majority of papers focus on operational decisions, followed by on tactical decisions, but few studies are available on strategic decisions, which is a broad area for future work.

Deterministic mathematical modeling is a part of optimization that has been broadly employed with the aim of decision making in real-world problems. In general, optimization involves finding the best solution for an objective function by limiting the search to specific conditions and constraints. The deterministic approach assumes that the data and parameters are known and have been used in many applications such as scheduling; however, in the presence of uncertainty (variable processing times) in a system, it may not give a realistic solution. Moreover, the presence of this uncertainty can make the optimal solution of a deterministic model infeasible or sub-optimal to the decision making problem. As a result, the stochastic approach tries to find solutions that optimize a performance measure under the assumption that uncertain parameters are random variables with known distributions. In stochastic programming, some distributional property of the objective function is usually adopted as a criterion to compare performances metrics in the problem. In other words, stochastic programming is another name for the research of optimal decision making under uncertainty. The term "stochastic programming" accentuates a connection to mathematical programming and algorithmic optimization schemes. These considerations in stochastic programming prevail over other fields of study and distinguish stochastic programming from other fields.

Operations research historically focused on deterministic models, because it has some properties such as: simplicity and better computational tractability, readily available commercial/open-source software, avoiding effort needed in characterizing uncertainties for stochastic programming.

However, the solution of deterministic models might be compromised due to poor representation of real-world complexities.

Stochastic programming has many applications in real-world problems such as manufacturing (supply chain planning), transportation (airline scheduling), telecommunications (network design), electricity power generation (power adequacy planning), health care (patient & resource scheduling), agriculture (farm planning under weather uncertainty), forestry (wildfire emergency response planning), finance (portfolio optimization). Airline planning is one of the first applications of stochastic programming to find the best way to allocate aircraft routes to improve passenger service (Ferguson & Dantzig 1956). (Birge & Louveaux 1997) offer many examples to illustrate various aspects of stochastic programming models in terms of the number of stages, continuous or discrete variables, probabilistic constraints, and linear/nonlinear constraint and objective functions. Moreover, (Sarin et al. 2014) reported various approaches such as robust scheduling, reactive scheduling, fuzzy scheduling, and stochastic scheduling that have been developed to address uncertainty in scheduling. For further information we refer the reader to (Daniels et al. 1995), (Kouvelis et al. 2000), (Sabuncuoglu & Bayiz 2000), (Balasubramanian & Grossmann 2003), and (Sarin et al. 2014) for each category in order and a complete survey on decision making under uncertainty by (Krokhmal et al. 2011).

According to (Birge & Louveaux 1997) we can categorize random events and random variables in two major classes. In the first class, we place uncertainties that recur frequently on a short-term basis. For instance, uncertainty may happen to daily or weekly demands. This results in a model where capacity allocation cannot be adjusted every time period. As a result, it follows that the expectation in the second-stage represents a mean over the possible values of the random variables, of which many will occur. In the second class, we place uncertainties that can be indicated as scenarios, of which fundamentally only one or a small number are realized. This would be the issue in long-term models

where scenarios demonstrate the general trend of the variables. In the second-stage, only one scenario is realized (among all scenarios over which the expectation is taken).

A two-stage stochastic programming approach is one of the most common methods in appointment scheduling. (Berg et al. 2014), (Erdogan & Denton 2013), (Qu et al. 2013), (Muthuraman & Lawley 2008) and (Erdogan et al. 2015) have formulated two-stage stochastic programming models. For complete review of these literature we refer the reader to chapter two of this dissertation.

Most of the recent literature has applied risk-neutral two-stage stochastic programming, which is a traditional method that has been used in many studies we mentioned earlier. There is a variety on choosing objective functions. A commonly used criterion is the expected value, which can be regarded as the long-run average performance of a schedule. This method finds the expected value of the performance measure such as patient flow metrics in the objective function as the preference criterion. For example (Erdogan et al. 2015) in outpatient appointment scheduling and (Skutella & Uetz 2005) in machine scheduling problems have used expected value as a performance measure. (Daniels et al. 1995) indicate that a critical disadvantage of using the expected value as a performance measure is that it does not account for the risk-averse attitude of a decision maker. As a result, some researchers have focused on considering a risk measure to model formulation. For example, (Sarin et al. 2014) use CVaR as a criterion in the machine scheduling problem considering uncertainty in the system, and (De et al. 1992) use variance as a risk measure to determine expectation-variance based efficient schedules. We also formulate a risk-averse two-stage stochastic programming in chapter three, which is related to mean-risk objectives and can be used instead of risk-neutral objectives. They consider the effect of variability and specify the preference relations among the random variables using risk measures such as Conditional-Value-at-Risk (CVaR). A few optimization studies have proposed risk-averse objectives, such as the Markowitz mean-variance method (Mak et al. 2015) and (Qu et al. 2012)

and the Von Neumann–Morgenstern expected utility method (e.g., (Kemper et al. 2014); (Kuiper & Mandjes 2015); (LaGanga & Lawrence 2012) and (Vink et al. 2015).

In the second chapter of this dissertation we mainly focus on risk-neutral two-stage stochastic programming where the objective function considers the expected value as a performance criterion, and in the third chapter, we expand the model formulation to mean-risk two-stage stochastic programming in which we investigate the effect of considering a risk measure in the model. We apply Conditional-Value-at-Risk (CVaR) as a risk measure for the two-stage stochastic programming model.

The goal of this dissertation is designed as follows: first, patient scheduling, where we optimize weekly scheduling template for individual providers to improve patient satisfaction by minimizing direct and indirect wait times as well as balance workloads, and new patient assignments. Next, the framework for dynamically scheduling patients using scheduling template which allows operationalization of scheduling template while allows the possibility of scheduling multiple appointments at once. Second, robust scheduling through Conditional-Value-at-Risk (CVaR). We develop a risk averse approach to capture the effects of variability of random outcomes under certain realizations of the random data. While improving metrics on average, we ensure no subset of patients are experiencing extreme waiting times.

CHAPTER 2: RISK-NEUTRAL TWO-STAGE STOCHASTIC PROGRAMMING MODEL TO OPTIMIZE THE PATIENT FLOW METRICS AT OUTPATIENT CLINICS

2.1. Introduction

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, and much research has been done recently to apply different methodologies such as overbooking and designing optimized appointment scheduling systems to overcome these deficiencies (Erdogan & Denton 2013), (Zacharias & Pinedo 2014), (Muthuraman & Lawley 2008). 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, which result in better reimbursement rates by payers for providers (Feldman et al. 2014), (Rau 2011). Many studies have documented the no-show rate in medical practice. (Macharia et al. 1992) reported a 42% average no-show rate which ranges from 6% to 92% in outpatient clinics. (Berg et al. 2014) reported 13% to 24% no-show rates at endoscopy clinics for different service types. (Festinger et al. 2002) shows post intervention no-show rates ranging from 28% to 45%. (Dreiher et al. 2008) results show the overall no-show rate at OBGYN clinics as 30.1%. They investigated the strong relationship between patients' appointment delays and no-show cases in OBGYN specialty clinics. In psychotherapy appointments, a 21% no-show rate was reported by (DeFife et al. 2010).

Another patient flow metric is patients' appointment delays known as indirect wait time in the literature. (Hawkins & Irving 2017) conducted a survey 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 as well as the rates of physician Medicaid and Medicare acceptance in these areas.(Hawkins &

Irving 2017) did this survey in 2004, 2009, 2014, and 2017 and the results showed an increase in the indirect wait time in 2017 comparing to other years. In 2004, the statistics was reported for 15 mid-sized metropolitan markets between 88,000 and 143,000 people including 1414 medical offices in large metro markets and 494 medical offices in mid-sized metro markets. They reported the indirect wait time for cardiology, dermatology, obstetrics-gynecology, orthopedic surgery and family medicine, which we depicted average indirect wait time of obstetrics-gynecology clinic as it is the focus of this research in Figure 1. Table 1 provides average obstetrics-Gynecology indirect wait time in major markets: Atlanta, Boston, Dallas, Denver, Detroit, Houston, Los Angeles, Miami, New York, Philadelphia, Portland, San Diego, Seattle, and Washington, D.C. are reported.

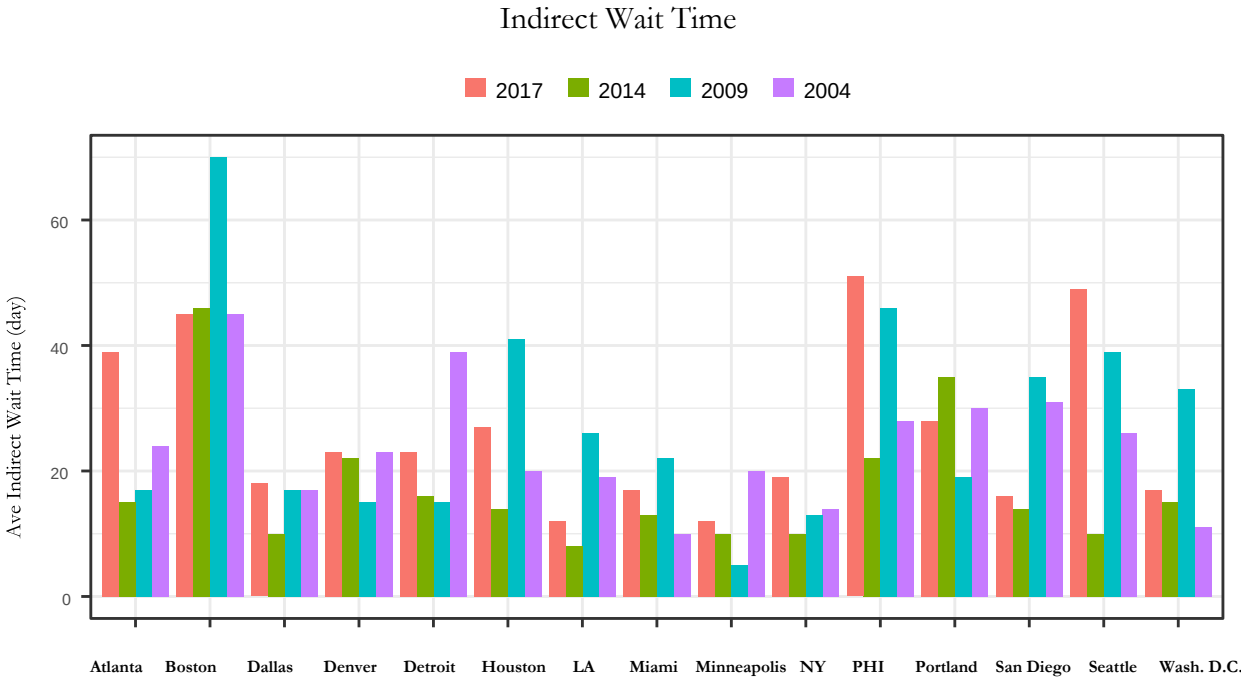

Fig. 1. Indirect wait time (day) in OBGYN clinic reported by (Hawkins & Irving 2017) for 15 cities in the United States

Average Obstetrics-Gynecology Appointment Wait Times, Major Markets*	
YEAR	DAYS
2017	26.4
2014	17.3
2009	27.5
2004	23.3

Table. 1. Indirect wait time (day) in OBGYN clinic reported by (Hawkins & Irving 2017) for 15 cities in the United States

In this chapter, we focus on the sources of inconsistencies such as no-show behavior, long direct and indirect wait time. The goal is to develop models that improve patient flow metrics: direct wait time (clinic wait time), indirect wait time considering patient's no-show behavior, stochastic server, follow-up surgery appointments, and overbookings. We develop a model for two purposes: 1) Patient Channeling, which means characterizing services rendered by the outpatient clinic and the individual physicians/staff within to channel new patients to the most appropriate service providers and address the needs of any clinical trials being supported by the providers; and 2) Patient Scheduling, the objective of which is to schedule both new and established patients for individual providers and facility locations while increasing throughput per session while providing timely care (e.g., minimizing the "indirect" wait-time between appointment desired date and appointment date), continuity of care, and overall patient satisfaction, as well as equity of resource utilization. This objective results in developing two models: 1) a method to optimize the (weekly) scheduling pattern for individual providers that would be updated at regular intervals (e.g., quarterly or annually) based on the type and mix of services rendered and 2) a method for dynamically scheduling patients using the weekly

scheduling pattern. Scheduling will entertain the possibility of arranging multiple appointments at once (e.g., both surgery and post-surgery follow-up visits can be scheduled together for improved care).

We introduce definitions and terms which will be used in this research. Some of them are from us and some from outpatient scheduling papers summarized in a survey by (Ahmadi-Javid et al. 2017).

Definitions and Terms:

- Appointment interval (slot): The time window between two consecutive appointment times.
- Appointment time: The start time of scheduled appointment for an individual patient.
- Block: Group of patients scheduled for the same appointment slot.
- Block size: The number of patients scheduled for the same appointment slot.
- Patient preference: A situation where a patient decides whether to accept the offered appointment time from call center or not; in other words, a patient accepts the appointment time with respect to his/her preference.
- Direct waiting time (clinic waiting time): The aggregate waiting time a patient experienced between the arrival to and exit from an individual server in the clinic. (Our research considers multiple servers).
- Indirect waiting time (delay): The time between the appointment request and the scheduled appointment time (Zacharias & Armony 2016).
- Flow time: The total time a patient spends in the clinic center (Cayirli & Veral 2003).
- No-show patient: A patient who does not show up for his\her appointment.
- No-show rate: The probability that the patient is a no-show case.
- Outpatient Appointment System (OAS): A main stream in an outpatient clinic that designs an appointment scheduling system with the aim of timely and convenient delivery and access to healthcare services for all patients (Gupta & Denton 2008).

- Outpatient clinic: A healthcare system that provides treatment and care to patients without an overnight stay in a health facility.
- Overtime: The difference between the available length of time session in a day for health services and the actual end of the service for the final patient in a clinic (Cayirli & Veral 2003).
- Panel size: The potential number of patients assigned to providers for services.
- Same-day appointment: An appointment that is scheduled on the same day that the patient asks for an appointment.
- Server idle time: The part of the consultation session that the server (or physician) is idle due to lack of patient(s).
- Service duration: The length of time a single patient spends with the service provider.
- Scheduled patient: A patient who makes an appointment before arriving at the clinic.
- Call center: An office in the medical service which provides appointment time to the individual patient. This center uses appointment scheduling template as a guidance to assign the appointment to each patient.

This chapter is organized as follows. Section 2.2 reviews the relevant literature. Section 2.3 describes the problem. Section 2.4 formulates a Two- Stage Stochastic Mixed-Integer Linear Program Model (two-stage SMILP). Solution of the two-stage SMILP provides the optimal capacity assigned for each time slot. Section 2.5 explains a demand generation simulation. In section 2.6, we introduce a dynamic appointment scheduling policy for actual appointment assignment for different patient types. Section 2.7 explains clinic simulations and direct wait time**.** In this section we calculate the direct wait time experienced by individual providers. Section 2.8 describes the case study and data driven from literature. Section 2.9 provides results and concluding remarks.

2.2. Literature review

In recent years, appointment scheduling in outpatient clinics has attracted much attention in health care delivery systems. Designing an effective appointment scheduling system in outpatient clinics results in a smooth flow of patients and work as well as consideration of patients' and physicians' preferences while matching supply and demand. As stated by (Gupta & Denton 2008) less attention goes to the benefit of patients compared to that of clinic services and physicians. Therefore, improving the performance of the health system with the aim of patient satisfaction that can be achieved by timely and appropriate health care delivery is the goal of a well-organized and reliable outpatient appointment scheduling system. Specifically, guaranteeing patients to get requested service with short time window as well as balancing the system's utilization, in order to prevent the system from over and under- utilization. On the other hand, matching demand and supply in the presence of uncertainty in the system is another issue. One solution can be taking care of enough inventories in production systems; however, service systems such as clinics, repair shops, airport transportations, manage request through appointments, (Liu et al. 2010). Moreover, there are a variety of uncertainties in the service systems, such as patient no-shows and patient cancellation which will affect system's performance. (Liu et al. 2010) developed a framework to find the possibility that patients may cancel or no-show at their time of appointments. There are many kinds of literature in outpatient appointment scheduling (AS). (Gupta & Denton 2008) discussed a variety of methods in modeling, optimization, and future work in appointment scheduling.

We review some of the categories in this section: static versus dynamic with solution methodologies, risk-neutral two-stage stochastic programming, clinic environment: multi versus single service. In some categories, we captured literature involved in direct and indirect waiting time, patient no-show and patient cancellation behavior as well as overbooking models.

2.2.1. Appointment scheduling problems

In survey papers by (Gupta & Denton 2008) and (Cayirli & Veral 2003), a complete review of the state of the art in modeling and optimization with future research studies is provided. (Cayirli & Veral 2003) divided decision making in outpatient appointment scheduling into static and dynamic models. (Cayirli & Veral 2003) and (Muthuraman & Lawley 2008) define static appointment scheduling when decisions about appointment times are made prior to the start of the appointment session while in dynamic case, appointment schedule may be modified later depending on the state of the system. In research by (Erdogan et al. 2015) static appointment scheduling is defined as a problem to find the optimal start times for a given number of patients to visit a stochastic server. In this case the number of patients is already known.

(Liu et al. 2010) developed a dynamic scheduling of outpatient appointment approach to assigning an appointment to each patient depending on the clinic's appointment schedule at the time of the patient's call. Comparing our research with papers in static and dynamic models, we schedule patients dynamically upon arrival of each request. Another category in our literature review is related to solution methodologies. The work by (Zeng et al. 2010), (Laganga & Lawrence 2007), and (Liu et al. 2010) used heuristics as solution methodology for appointment scheduling. Moreover, in another paper, (Zenios et al. 2000) used heuristic policy to allocate kidney to transplant patients dynamically. (Qu et al. 2013) in outpatient scheduling with a specialty for OBGYN apply Monte Carlo sampling based genetic algorithm to solve a mixed integer program. (Liu et al. 2010) also develop heuristic policy for dynamic appointment scheduling considering one patient type with no-show and cancellation behaviors to assign appointments to arrival calls on a daily pattern. Similarly, we develop a heuristic to assign an appointment to each arrival request.

Another classification on appointment scheduling is with respect to two waiting times: direct

and indirect. (Liu et al. 2010), (Erdogan & Denton 2013), (Qu et al. 2013), (Muthuraman & Lawley 2008) and (Zacharias & Armony 2016) are among the most recent studies on appointment scheduling which consider waiting time in model formulations. In a paper by (Zacharias & Armony 2016), direct waiting time/clinic delay is physical waiting time experienced by patients once they arrive at the clinic, and indirect waiting time/appointment delay is defined as the time window between the appointment request and the offered appointment. In the research of (Zacharias & Armony 2016), crucial characteristics such as the randomness of service time and patient punctuality as well as patient no-show behavior are addressed. Moreover, the optimal number of appointment slots per day and the size of the medical practice panel are captured. In the problem formulation, both direct and indirect waiting times are addressed; next, based on the diffusion approximations technique, they end up with a closed form formulation that includes a performance measure of maximizing the long-run average daily net profit of a medical system while providing care to patients. Similar to the research by (Zacharias & Armony 2016), we minimize indirect and direct waiting time in the model.

In another study by (Qu et al. 2013), a weekly schedule pattern in outpatient clinics for an OB-GYN specialty considering different service types and different providers is found. They develop a model formulation in two phases; in the first phase they formulate a mixed integer program and capture the scheduling pattern, and in the second phase, they propose a stochastic mixed integer program to assign appointment start times while minimizing patient direct waiting times and provider idle/over time. A Monte Carlo sampling based genetic algorithm is developed to solve the two-stage mixed integer program. Similar to the research by (Qu et al. 2013), we get the appointment schedule, schedule patients dynamically and monitor direct waiting time; however, we also capture the scheduling pattern after solving a two-stage stochastic mixed integer linear program and minimize the indirect waiting time in our model formulation.

Many articles as discussed above captured direct waiting time in their model formulation direc-

tly, but few researchers have addressed indirect waiting time. One of the contributions of our work is the control of indirect waiting time of the system in the model formulation while monitoring the direct waiting time indirectly as a feedback process which is explained in section (2.7).

2.2.2. Two-Stage Stochastic Programming in Appointment Scheduling

A two-stage stochastic integer programming approach is one of the most common methods in appointment scheduling. (Berg et al. 2014), (Erdogan & Denton 2013), (Qu et al. 2013), (Muthuraman & Lawley 2008) and (Erdogan et al. 2015) formulated two-stage stochastic integer programming models. (Berg et al. 2014) presented optimal booking methods in outpatient clinics. They employed 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. The number of appointments reserved for a given day, the relationship between the number of reserved patients and the likelihood of nonattendance, the optimal priority of patients during the day, as well as the optimal arrival model and whether it is optimal to consider double booking in case of cancellation or no-show cases is investigated. However, arrival delay and rescheduling in a given day are not allowed in the model.

(Erdogan & Denton 2013) formulated the appointment scheduling in two model levels: first is a two-stage stochastic linear program (2-SLP) for static appointment scheduling capturing no-show behavior; second, dynamic scheduling is formulated by a multistage stochastic linear program (M-SLP). The authors used a decomposition algorithm, and computational experiments are reported.
In research by (Muthuraman & Lawley 2008), the stochastic overbooking model in outpatient appointment scheduling for clinical use is modeled. Patient waiting time, provider over time and patient revenue are considered in the objective function of the model formulation.

The literature on appointment scheduling we discussed has limitations: there is no simultaneo-

us consideration of direct and indirect wait times along with providers' workloads. Mostly risk-neutral approaches and limited planning horizon are considered.

In this research, we devise a two-stage stochastic mixed integer program for appointment scheduling and consider demand uncertainty in the system that takes into account the no-show behavior of patients. We also assume that assignments cannot be changed once the appointment is scheduled for the patient. Moreover, we are interested in determining the appointment time-slot in a service session when the appointment should be scheduled. In addition, we open free time-slots in our model formulation for emergency/post-surgery follow-up arrivals and calculate the direct waiting time of the system simulating the clinic with multiple servers and control direct waiting time indirectly.

2.3 Problem description

We design an appointment scheduling model capturing multi-type patient channeling to different provider levels in the OBGYN clinic specialty. The objective is to improve patient flow through outpatient clinics using efficient appointment scheduling policies. We improve indirect waiting time in our formulation settings as part of the objective function, and direct waiting time at the clinic specialty as part of our constraints in our model. Direct waiting time/clinic delay, is physical waiting time experienced by the patients once they arrive at the clinic and indirect waiting time/appointment delay is defined as the time window between the appointment request and the offered appointment, (Zacharias & Armony 2016). The objective of the decision-making problem in the first-stage is to balance a provider's workload between different clinic sessions as well as among each time slot. The provider's workload is controlled by channeling individual patient's type to the appropriate provider in the constraints of the model during each work day. Based on the research on OBGYN specialty clinics by (Qu et al. 2013) and (Lenin et al. 2015), we divide patients into three categories and, consequently, seven patient types with respect to the expected service time duration for each patient type, Table.2.

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

Table.2. OBGYN patient types

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) with identical service slots of 15 minutes, which is common in practice. Moreover, as many providers are different in their practice styles in specialty clinics, the model opens free capacity for lunch hours, office work for providers and in some cases appointments for follow-up surgery. Service time duration for each patient type is based on the literature on OBGYN clinic (Qu et al. 2013) and (Lenin et al. 2015). In the research by (Lenin et al. 2015), data are collected from the West Little Rock (WLR) clinic operated under the University of Arkansas for Medical Sciences (UAM). The research framework is shown in Figure 2. First, the risk-neutral two-stage stochastic programming model employs the input data including supply and demand parameters and produces a weekly scheduling template. This scheduling template specifies appointment allocation of patient's requests considering patient types and resource availability to different time slots. Next, in clinic simulation we evaluate the performance of the patient flow of the weekly scheduling template. If the flow of the patients for a week are not satisfactory, additional constraints are added to the model to avoid the sequence causing this unpleasant condition within the optimal template and we re-optimize

the model. We will continue this process until we obtain the optimized weekly scheduling. In the next step, the optimal scheduling template is ready to be used with the call center for actual appointment assignment. This scheduling template is used as a guide for a whole planning horizon. After assigning the appointments, we check the true patient flow for the clinic and see if the que of the patients is satisfactory or not. We continue this process for the planning horizon until we receive a patient flow, which is unsatisfactory. At this time, we re-optimize the scheduling template.

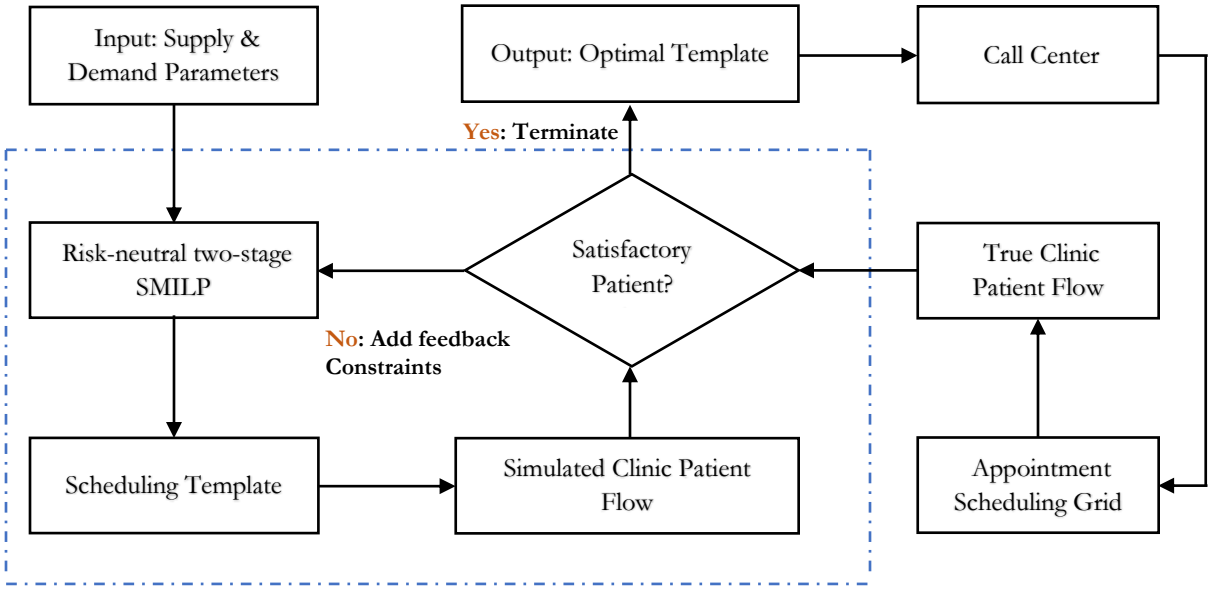

Fig. 2. Research framework

2.4 Two- Stage Stochastic Mixed-Integer Linear Program Model (two-stage SMILP)

Two-stage stochastic programming methodology is a mainstream technique in model formulation under uncertainty and inexactness in data. Decisions without complete information on random events are called first-stage decisions. Soon thereafter, full information is received on the realization of some random vector, and the second-stage data become known; then the second-stage decision is made. This chapter addresses Two-Stage Stochastic Mixed-Integer Linear Program (two-stage SMIPL) model, where the first stage consists of decisions on the number of capacities for the scheduling template and some penalty costs for over/under utilization of time slots, and the second stage involves

some recourse, such as a penalty for indirect wait time as well as capacity violation cost. A generic formulation of this class of problems is

$$z^* = min_{x\epsilon X} c^T x + \mathbb{E}_{\Omega}[f(x, \xi(\tilde{\omega}))],$$

where

$$f(x, \xi(\omega)) = min_{y \geq 0}\{q(\omega)^T y \mid D_y \geq h(\omega) - T(\omega)x\},$$

x denotes the first-stage appointment capacity decision, X denotes the first-stage feasible set involving constraints to control critical factors for patients and providers for scheduling purposes, $\omega\epsilon\Omega$ denotes a scenario that is unknown when the first-stage decision x has to be made, but that is known when the second-stage recourse decision y is made, Ω is the set of all scenarios, and c denotes the penalty cost for over/under utilization for new patients as well as each time-slot. We assume that the probability distribution *P* on Ω is known in the first stage. The quantity $f(x, \xi(\tilde{\omega}))$ represents the optimal value of the second-stage recourse problem corresponding to the first-stage x and the parameters $\xi(\omega) = (q(\omega), h(\omega), T(\omega))$. In the following subsections, we first introduce model formulation in section (2.4.1). Solution scheme: sample average approximation (SAA) in section (2.4.2). Then, we explain demand generation addressing uncertainty in the system in section (2.5).

2.4.1 Model formulation

The objective of the decision-making problem in the first-stage balances a provider's workload not only among morning/afternoon sessions, but also in each time-slot of the clinic. In our model formulation, the first-stage determines the amount of capacity reserved for each patient type assigned to each provider for individual time-slots in a weekly pattern which will be used for a whole month. In the second-stage, we determine time-slot utilization for individual patient types assigned to each provider for individual time-slots under scenario ω. We use notations denoted in Table 3 for the model formulation.

Set	
T	Set of planning horizon
$\mathcal{R}$	Set of providers
$\mathcal{N}$	Set of patient types
N'	Set of new patient type
Ω	Set of all scenarios
RPt	Set of risk factors for different patient type
RPr	Set of risk factors for different provider levels
$\mathcal{H}$	Set of free time slots for each provider over time horizon T
$\mathcal{S}$	Set of morning/afternoon sessions over time horizon T
η	Set of feedback sequence over morning/afternoon session of every day
β	Set of patients scheduled for specific clinic day
ξ	Set of exam rooms in the clinic
Γ	Set of call, desired and appointment times, indexed by $\gamma(t)\epsilon\Gamma$ containing time-slot, $t\epsilon T$
Parameter	
a_j	Number of new patients desired by provider, $j\epsilon\ \mathcal{R}$
cf_i	Risk factor for patient type, $i\epsilon\mathcal{N}$
CF_j	Risk factor for provider, $j\epsilon\ \mathcal{R}$
tlr_j	Tolerance factor of provider, $j\epsilon\ \mathcal{R}$
Δ_j	Cost of additional capacity of provider, $j\epsilon\ \mathcal{R}$
ρ_j	Cost of new patient type for provider, $j\epsilon\ \mathcal{R}$
c_j	Free capacity for provider, $j\epsilon\ \mathcal{R}$ over time horizon T
p_i	Average no-show probability of patient type, $i\epsilon\mathcal{N}$
M	A large number
$\mathcal{G}$	Number of time-slots per week
$\lvert\mathcal{S}\rvert$	Cardinality of $\mathcal{S}$
λ	Penalty parameter for penalty variable for each time-slot, $t\epsilon T$
$d_{i,j,\gamma(t)}(\omega)$	Demand of patient-type, $i\epsilon\mathcal{N}$ ask for provider, $j\epsilon\ \mathcal{R}$, with call and desired time set $\gamma(t)\epsilon\Gamma$ under scenario, $\omega\epsilon\Omega$
First-stage decision variables	
$x_{i,j,t}$	Number of patient type, $i\epsilon\mathcal{N}$ assigned to provider, $j\epsilon\mathcal{R}$ per time-slot, $t\epsilon T$
e_j	Penalty variable for provider, $j\epsilon\ \mathcal{R}$ w.r.t. new patient type
$z_{j,t}$	1 if time-slot, $t\epsilon T$ is free for provider, $j\epsilon\ \mathcal{R}$, else 0
dev_t	Penalty variable for each time-slot, $t\epsilon T$
Second-stage decision variables	
$y_{i,j,\gamma(t)}(\omega)$	Time slot utilization for number of type, $i\epsilon\mathcal{N}$ patient asked for provider, $j\epsilon\ \mathcal{R}$ with call, desired and appointment time set $\gamma(t)\epsilon\Gamma$ under scenario, $\omega\epsilon\Omega$
$b_{j,t}\ (\omega)$	Capacity slack variable for provider, $j\epsilon\ \mathcal{R}$, time-slot, $t\epsilon T$, under scenario, $\omega\epsilon\Omega$

Table. 3. Notation used in Risk-neutral two-stage SMILP model

First-stage objective function:

$$\min \sum_{j \in \mathcal{R}} \rho_j \cdot e_j + \lambda \cdot \sum_{t \epsilon T} dev_t + \mathbb{E}_{\Omega}[f(x, \xi(\tilde{\omega}))] \qquad (P)$$

First-stage constraints:

$$e_j + \sum_{t \epsilon T} \sum_{i \epsilon N'} x_{i,j,t} \geq a_j \qquad N' \subset N, \forall j \in \mathcal{R} \ (1)$$

$$e_j - \sum_{t \epsilon T} \sum_{i \epsilon N'} x_{i,j,t} \geq -a_j \qquad N' \subset N, \forall j \in \mathcal{R} \ (2)$$

$$dev_t - \sum_{j \epsilon \mathcal{R}} \sum_{i \epsilon N} x_{i,j,t} + \sum_{i \epsilon N} \sum_{j \epsilon \mathcal{R}} \sum_{t \epsilon T} x_{i,j,t} / \mathcal{g} \geq 0 \qquad \forall t \in T \ (3)$$

$$dev_t + \sum_{j \epsilon \mathcal{R}} \sum_{i \epsilon N} x_{i,j,t} - \sum_{i \epsilon N} \sum_{j \epsilon \mathcal{R}} \sum_{t \epsilon T} x_{i,j,t} / \mathcal{g} \geq 0 \qquad \forall t \in T \ (4)$$

$$\sum_{i \epsilon N} cf_i \cdot x_{i,j,t} \leq CF_j \qquad \forall t \in T, j \in \mathcal{R} \ (5)$$

$$\sum_{t \in \mathcal{S}} \sum_{i \epsilon N} cf_i \cdot x_{i,j,t} \leq |\mathcal{S}| CF_j - tlr_j \qquad \mathcal{S} \subset T, \forall j \in \mathcal{R} \ (6)$$

$$\sum_{t \epsilon \mathcal{H}} z_{j,t} = c_j \qquad \mathcal{H} \subset T, \forall j \in \mathcal{R} \ (7)$$

$$\sum_{i \epsilon N} x_{i,j,t} \leq \mathrm{M} \cdot (1 - z_{j,t}) \qquad \forall t \in \mathcal{H} \subset T, j \in \mathcal{R} \ (8)$$

$$\sum_{i,j,t \epsilon \eta} x_{i,j,t} \leq |\eta| - 1 \qquad \eta \subset \beta, \eta \neq \emptyset \ (9)$$

$$x_{i,j,t} \in \mathbb{Z}^+, e_j \in \mathbb{R}^+, z_{j,t} \in \{0,1\}, dev_t \in \mathbb{Z}^+, \mathcal{g} \in \mathcal{G} \qquad (10)$$

Second-stage objective function:

$$f(x, \xi(\omega)) =$$

$$\min \sum_{i \epsilon N} \sum_{j \in \mathcal{R}} \sum_{\gamma(t) \epsilon \Gamma} y_{i,j,\gamma(t)}(\omega) \cdot d_{i,j,\gamma(t)}(\omega) \cdot \vartheta + \sum_{j \in \mathcal{R}} \sum_{\gamma(t)/\{tc,td\} \epsilon \Gamma} b_{j,\gamma(t)}(\omega) \cdot \Delta_j$$

Second-stage constraints:

$$\sum_{\gamma(t)/\{ta\} \epsilon \Gamma} (1 - p_i) \cdot y_{i,j,\gamma(t)}(\omega) \cdot d_{i,j,\gamma(t)}(\omega) \leq x_{i,j,t} + b_{j,t}(\omega) \qquad \forall i \in N, j \in \mathcal{R}, t \in T \ (11)$$

$$\sum_{\gamma(t)/\{tc,td\} \epsilon \Gamma} y_{i,j,\gamma(t)}(\omega) = 1 \qquad \forall i \in N, j \in \mathcal{R}, \gamma(t)/\{ta\} \in \Gamma \ (12)$$

$$0 \leq y_{i,j,\gamma(t)}(\omega) \leq 1, \ b_{j,t}(\omega) \epsilon \mathbb{R}, \omega \epsilon \Omega \qquad (13)$$

In the above formulation, constraints (1) and (2) check the difference between the desired nu-

mber of new patients by individual providers and the assigned number of new patients to each provider. In other words, the equity of new patients among all providers is being evaluated by constraints (1) and (2). Constraint (3) and (4) calculate all capacities reserved for each time-slot and find average of capacities reserved over the week. Finally, they find the deviation between capacities reserved for each time-slot and average the amount over the week. Next, this deviation is penalized in the objective function (P). In constraint (5), provider workload in each time slot of the clinic is controlled, and individual patient type is channeled to each provider. However, constraint (6) is to balance the provider workload among clinic sessions while channeling patient types to the providers. Constraint (7) opens free capacity for each provider based on the desired number of time slots by individual providers through afternoon sessions. These free capacities are reserved for emergency/ post-surgery follow-up appointment requests. Constraint (8) guarantees there will be no assignments in time slots obtained by constraint (5). Constraint (9), which is called the feedback *constraint*, is to remove the sequence of patients whose violated clinic wait time threshold. In the second-stage, constraint (11) doesn't allow each time-slot's utilization to exceed the capacity reserved in the first-stage mixed-integer linear problem. In the second-stage, capacities are determined based on first-stage decisions.

Finally, constraint (12) assigns appointment time to each demand arrival. The objective function (P) in two-stage mixed-integer linear problem penalizes the system's over/under utilization in terms of time slot. In the first part of the objective function, the model penalizes the over/under utilization of time-slots reserved for new patient types for an individual provider, and in the second part of the objective function, indirect waiting time (the time between a patient's desired time and the assigned appointment time) in terms of time slot is penalized. In the second-stage objective function, ϑ denotes $f(ta - tc).(ta - td)$, where $f(ta - tc) = (ta - tc)^{-\frac{1}{2}}$ is called the penalty function (super-linear function) and controls the indirect waiting time of the system; This function considers

fairness in assigning appointment with delays to patient requests. tc and ta denote call and appointment times, respectively.

2.4.2 Solution Scheme: Sample average approximation (SAA)

Referring to (Verweij et al. 2003), the sample average approximation (SAA) method is an approach to solve stochastic optimization problems. Moreover, sample average approximation approach brings some advantages to two-stage stochastic programming. Firstly, the two-stage SMILP off the shelf solvers can typically solve instances with few number of scenarios. However, a typical problem instance in a practical case would have thousands of scenarios. Sample average approximation (SAA) method is a method to handle this problem. Approximating an optimal solution of stochastic programming with small number of scenarios results in monotonically better solution when we increase the number of scenarios. Secondly, SAA is useful when the number of scenarios is unknown.

We use the sample average approximation (SAA) to reduce to the size of the problem by repeatedly solving it with a smaller set of scenarios. We generate random samples with $\mathcal{N} < |\Omega|$ realizations of the uncertain parameters and approximate the expected recourse costs by the sample average function $\frac{1}{\mathcal{N}}\sum_{n=1}^{\mathcal{N}} f(x, \tilde{\omega})$.

As a result the problem (1) – (13) is approximated by the following SAA problem:

$$\min \sum_{j \in \mathcal{R}} \rho_j \cdot e_j + \lambda \cdot \sum_{t \epsilon T} dev_t + \frac{1}{\mathcal{N}} \sum_{n=1}^{\mathcal{N}} f(x, \omega^n) \qquad (14)$$

If we solve the SAA problem (14) with independent samples for many times, the outcome can be more efficient than increasing the sample size $\mathcal{N}$. For complete procedure we refer the readers to (Schutz et al. 2009) and (Santoso et al. 2005); however, we include it here for complementary:

1. Generate $\mathcal{M}$ independent samples of size $\mathcal{N}$ and solve the SAA problem in below:

   $$\min \sum_{j \in \mathcal{R}} \rho_j \cdot e_j + \lambda \cdot \sum_{t \epsilon T} dev_t + \frac{1}{\mathcal{N}} \sum_{n=1}^{\mathcal{N}} f(x, \omega^n)$$

2. Calculate the average of all optimal objective function values from the SAA problems:

$$\bar{v}_{\mathcal{N},\mathcal{M}} = \frac{1}{\mathcal{M}} \sum_{m=1}^{\mathcal{M}} v_{\mathcal{N}}^{m}$$

$$\delta^2_{\bar{v}_{\mathcal{N},\mathcal{M}}} = \frac{1}{(\mathcal{M}-1)\,\mathcal{M}} \sum_{m=1}^{\mathcal{M}} (v_{\mathcal{N}}^{m} - \bar{v}_{\mathcal{N},\mathcal{M}})^2$$

where $v_{\mathcal{N}}^{m}$ is the optimal objective function value, $\bar{v}_{\mathcal{N},\mathcal{M}}$ the average objective function value denotes a statistical lower bound on the optimal objective function value for the original problem (1)–(13) (Norkin et al. 1998), (Mak et al. 1999), and (Verweij et al. 2003).

3. Find a feasible first-stage solution $\bar{x}$ and estimate the objective function value of the original problem with sample size $\mathcal{N}'$ which is very larger than $\mathcal{N}$. $\mathcal{N}'$ is generated independently of the samples used in the SAA problems. Since the first-stage solution is fixed and this step involves the solution of the second-stage problems, we can choose $\mathcal{N}'$ larger than $\mathcal{N}$.

$$\hat{g}_{\mathcal{N}'}(\bar{x}) := \sum_{j \in \mathcal{R}} \rho_j \cdot e_j + \lambda \cdot \sum_{t \epsilon T} dev_t + \frac{1}{\mathcal{N}'} \sum_{n=1}^{\mathcal{N}'} f(x, \omega^n)$$

The estimator $\hat{g}_{\mathcal{N}'}(\bar{x})$ is an upper bound on the optimal objective function value. We can estimate the variance of $\hat{g}_{\mathcal{N}'}(\bar{x})$as follows:

$$\delta^2_{\mathcal{N}'}(\bar{x}) = \frac{1}{(\mathcal{N}'-1)\,\mathcal{N}'} \sum_{n=1}^{\mathcal{N}'} (\sum_{j \in \mathcal{R}} \rho_j \cdot e_j + \lambda \cdot \sum_{t \epsilon T} dev_t + f(x, \omega^n) - \hat{g}_{\mathcal{N}'}(\bar{x}))^2$$

4. Calculate the estimators for the optimality gap and its variance. Referring to steps 2 and 3, we obtain:

$$gap_{\mathcal{N},\mathcal{M},\mathcal{N}'}(\bar{x}) = \hat{g}_{\mathcal{N}'}(\bar{x}) - \bar{v}_{\mathcal{N},\mathcal{M}}$$

2.5 Demand Generation

Demand parameters are parts of inputs for two-stage SMILP model. We assume the number of patients asks for appointment are uncertain and generate demand for appointment requests for many scenarios. Demand is generated with respect to these scenarios: We assume six-month time horizon which patient calls arrive in the first four months and their desired time could be from when they call until the end of time horizon (six months). In demand generation, we assume the difference

between call day and desired day of a week (5 days) and the difference between call week and desired week follow distributions. Reviewing other literature, e.g. (Liu et al. 2010), (Patrick et al. 2008), and (Gupta & Denton 2008), we assume the daily patient arrival follows the Poisson distribution.

2.6 Dynamic Appointment Scheduling

After finding an optimal weekly appointment scheduling pattern from the two-stage SMILP model, the call center uses the solution from the two-stage SMILP on daily dynamic appointment assignment. This is referred to as *Call Center appointment assignment*. Next, we simulate the call center with demand generation and develop the heuristic policy to assign an appointment time to each patient arrival. Patients are quoted their appointment times upon requests for appointment. The sequence of appointments may change over time as the appointment schedule evolves; however, once an appointment time is assigned for a given patient, it cannot be changed. Our demand generation has these parameters: patient type, provider, call time, and desired time for one scenario. Upon arrival of each appointment request for a day, the appointment is offered with respect to the sorted max capacity in the first week from the patient's desired time. If an appointment is not accepted by the patient within the first week, the first available appointment time in the remaining month will be offered, then, if patient still doesn't accept the appointment in the first month, we offer the available time-slot in the remaining time window until the patient accepts the appointment time. We summarize the index heuristic policy below.

Index heuristic policy:

Input weekly appointment scheduling template *S*, demand set *D* for time horizon T,
and appointment acceptance threshold τ

1: **for** demand arrival *D* in day i:
2: $Capacity \leftarrow \{\}$
3: **for** $t \in \{DT, \dots, DT+T\}$:
4: find the corresponding capacity for time slot t, $I_t = x_t$, $Capacity = I_t$
where DT is patient's desired time, T is one-week time window,
5: **for** $j \in length\{Capacity\}$:
6: find $t^* = argmax(I_t), t \in \{DT, \dots, DT+T\}$and offer time slot t^* to the patient,
7: If τ meets, update S: $I_t = x_t - 1$ and go to step 1;
otherwise, go to step 8
8: **for** $t \in \{DT, \dots, DT+T'\}$:
9: search the first available slot, $I_t = x_t, x_t > 0, t \in \{DT, \dots, DT+T'\}$, where T'
is one-month time window
10: If τ meets, update S: $I_t = x_t - 1$ and go to step 1;
otherwise go to step 11;
11: **for** $t \in \{DT, \dots, DT+T''\}$:
12: assign appointment slot in the remaining time horizon T'', for $I_t = x_t, x_t > 0$, update
S: $I_t = x_t - 1$ and go to step 1.

Output: updated weekly appointment scheduling template S.

Note that the remaining time window threshold after the first week horizon depends on the patient type urgency. For some patients we may need to consider one month, whereas for other patient types this threshold could be in months. It depends on the patient service category.

2.7 Clinic simulation

As we discussed in the literature review section, most of the research done on outpatient clinics aims to minimize the direct waiting time of the clinic in the model formulation of two-stage mixed integer programming. However, we monitor the clinic waiting time of the system by simulating the clinic using the following formulation. We check the daily expected waiting time of the clinic for a sequence of patients for a given day. After each day, we check if the expected waiting time of the clinic

for the given day is greater than some threshold; we avoid creating such a sequence of patients in the future of the planning horizon by removing that sequence in the first-stage of the model formulation, using constraint (9). This approach will affect other flow metrics such as the system's over time and idle time. The clinic has multiple servers, and service times in each server are random variables. The objective is to minimize the expected patient waiting time, as well as the provider's overtime and idle time with respect to the established day length, $\mathcal{s}$. Figure 3 depicts the appointment start time and process time for scenario ω for a single server $\hbar$ in the clinic.

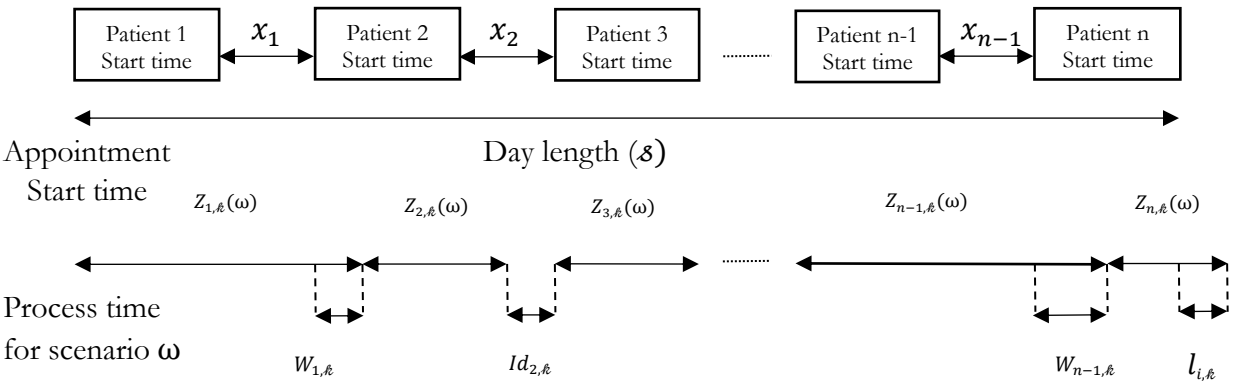

Fig. 3. Appointment start time and process time for scenario ω for a single server $\hbar$ in the clinic

We calculate patient waiting time $W_{i,\hbar}$ by developing formula that consider multiple servers in the system. Moreover, provider's over time and idle time can be calculated by $l_{i,\hbar}$ and $Id_{i,\hbar}$:

$$W_{1,\hbar} = 0 \quad ,\forall\, \hbar = 1, \dots, k$$
$$W_{i,\hbar} = (W_{i-1,\hbar} + Z_{i-1,\hbar} - x_{i-1,\hbar})^+ \quad , i = 2, \dots, n, \hbar = 1, \dots, k$$
$$Id_{i,\hbar} = (-W_{i-1,\hbar} - Z_{i-1,\hbar} + x_{i-1,\hbar})^+ \quad , i = 2, \dots, n, \hbar = 1, \dots, k$$
$$l_{i,\hbar} = (W_{n,\hbar} + Z_{n,\hbar} + \sum_{i=1}^{n-1} x_{i,,\hbar} - \mathcal{s})^+$$

where $Z_{i,\hbar}$ is the independent and identically distributed service duration for patient i at clinic room $\hbar$, $x_{i,\hbar}$ is customer allowance (inter-arrival time between patient i and $i+1$), $(.)^+$ indicates $\max(.,0)$ and d is session length. The total waiting time of the system for a given day equals $\sum_{i\epsilon\beta}\sum_{\hbar\epsilon\xi} W_{i,\hbar}$, where β is the set of patients scheduled for an individual clinic day, ξ is the set of clinic rooms in the

clinic, k is the number of clinic rooms, and n is the number of patients. The flow of patients at the clinic are shown in Figure 4.

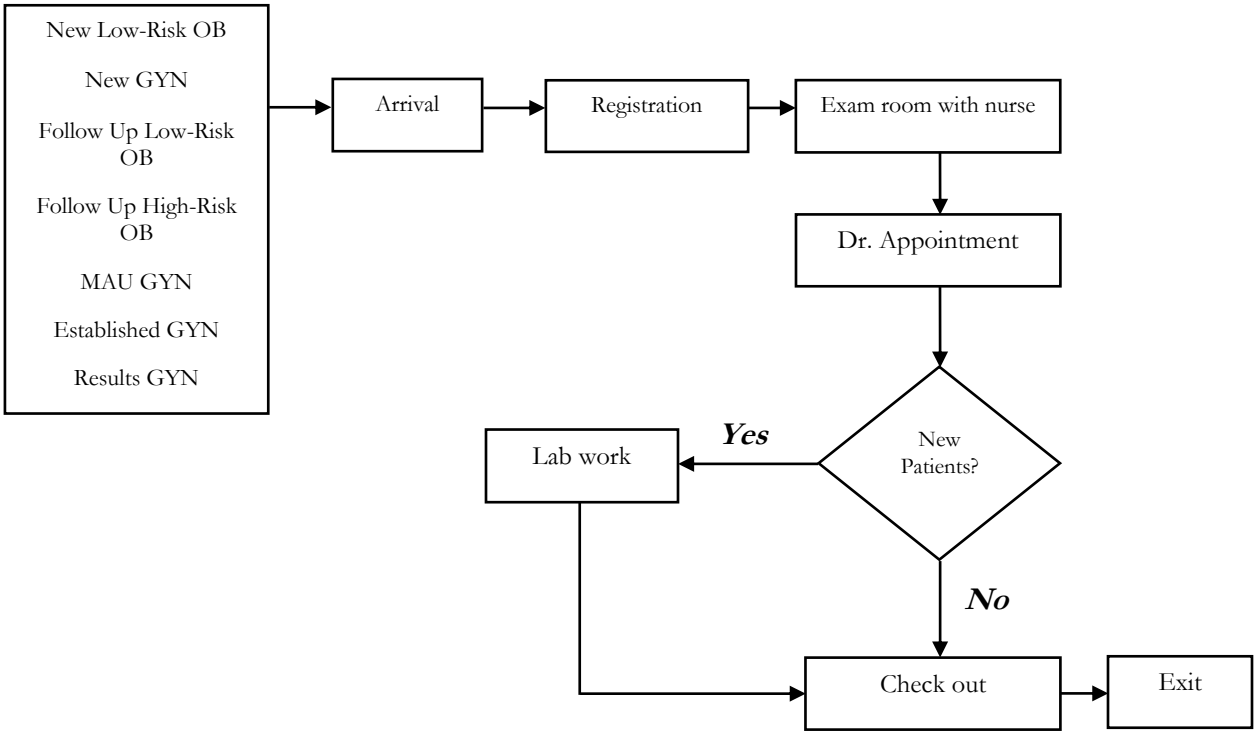

Fig. 4. Flow of OBGYN patients in the clinic

2.8 Case Study

In this section, we report a case study that demonstrates how well the proposed mean-risk two -stage SMILP model approach performs in terms of the multi-category outpatient appointment scheduling for the women's clinic studied. The clinic characteristics and patient demand data used in the case study are acquired from the literature of women's specialty clinics. The values of the parameters in the risk-neutral two-stage SMILP model are selected from (Qu et al. 2013) and (Lenin et al. 2015) as well as some from preliminary numerical experiments and are denoted in Table 5.

2.8.1 Data and study design

Studying the literature in OBGYN clinics, the common issue is related to the time, equipment and exam rooms scheduled for several service categories. Since each patient type needs specific services such as prenatal and follow-up care for routine pregnancy, high risk pregnancy, management of miscarriage in new and follow up cases with different exam equipment and resources, (Qu et al. 2013) divided patient types with respect to required service types into three categories. Consequently, there are seven patient types with respect to the expected service time duration for each patient type (Table 2). In this case study, each clinic session is defined as a day and is divided into 16 time slots with the identical service time 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 stations in the clinic such as time spent at registration, with a nurse or provider, lab work and check-out are included in the clinic simulation and taken from (Qu et al. 2013) and (Lenin et al. 2015). In the research by (Lenin et al. 2015), data are collected from the West Little Rock (WLR) clinic operated under the University of Arkansas for Medical Sciences (UAM). In the case study, we use two demand cases. In the first case, the average weekly number of demands is taken from (Qu et al. 2013), and as we expect increase in future demand, in order to estimate the scalability of the solutions, the demand was increased twofold. Table 4 shows the weekly demand cases with service time duration and the no-show rate. The proposed risk-neutral two-stage SMILP approach is used to determine weekly scheduling templated for these two cases.

Service Category	Service Type	Service time (minutes)			no-show rate	Avg. number of requests for service	
		Avg.	Std.	Distribution $LN(\mu, \delta^2)$		Case-1	Case-2
Low Risk OB	New Low-Risk OB	25	8	LN(3.17, 0.10)	0.162	4	8
	Follow Up Low-Risk OB	6	3	LN(1.68, 0.22)	0.053	22	44
High Risk OB	Follow Up High-Risk OB	10	6	LN(2.15, 0.31)	0.080	35	70
Gynecology	New GYN	18	12	LN(2.71, 0.37)	0.488	16	32
	MAU GYN	13	3	LN(2.54, 0.05)	0.487	4	8
	Established GYN	10	5	LN(2.19, 0.22)	0.384	17	34
	Results GYN	15	4	LN(2.67, 0.07)	0.321	5	10
Nurse	All service category			LL(100,2.92,417) / 60.0			

Table. 4. Weekly demand, no-show rate, and service time distribution for each patient types (time spent with provider and nurse)

We assume that there are 2 sessions: morning and afternoon, and each session has 8 time-slots. Based on the data driven from (Qu et al. 2013) in OBGYN clinics, services rendered for different patient types are considered in different sessions of a week day. Moreover, women's 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 depending on the availability of multiple providers in any clinic session. Since multiple providers are assigned for each day, overbooking is allowed for each time slot.

Notation	Description	Value
K	Total number of physicians available in each clinic session	2
N	Number of time slots in each clinic session	16
Δ_j	Cost of additional capacity of provider	[2000, 2000]
a_j	Number of new patients desired by provider	[10,10]
cf_i	Risk factor for patient type	[1.67, 0.4, 0.67, 1.2, 0.87, 0.67, 1]
CF_j	Risk factor for provider	[1.67, 1.67]
tlr_j	Tolerance factor of provider	[4.5, 4.5]
ρ_j	Cost of new patient type for provider	1.7
c_j	Free capacity for provider, $j \in \mathcal{R}$ over time horizon T	[2, 2]
M	A large number	4.8
$\mathcal{S}$	Set of morning/afternoon sessions over time horizon T	8
$\mathcal{g}$	Patient acceptance threshold for the first week	0.5 <= threshold< 1
$\mathfrak{I}$	Patient acceptance threshold for one month	0.2 <= threshold< 0.5
T	Time horizon	120 days
$\mathcal{f}$	Steady state	61– 100 days

Table. 5. Two-stage SMILP model setting parameters in the case study

2.9 Computational Results

The calculations were carried out on a Dell, 64-bit operating system, and 80 GB RAM. The solution scheme is implemented in Python 2.7.12. Gurobi is used as a solver for two-stage SMILP and SAA. In this section we present the significance of applying risk-neutral two-stage SMILP approach versus base-case scenario. We define base-case scenario with simulating clinic and call center using the same scenario as we design in risk-neutral two-stage SMILP approach. In the call center simulation, we consider corresponding risk factor for each patient type and each provider for each time-slot as well as each day sessions: morning and afternoon meaning the capacity of each time slot cannot exceed

the risk factor of individual providers and for the appointment scheduling we use the same index heuristic policy as we used in risk-neutral two-stage SMILP approach. Similarly, we use the same rules for clinic simulation and find direct waiting time and indirect waiting time. We consider three threshold levels by calculating 50%, 65%, and 80% quantiles of daily expected waiting time for two months' time horizon to check whether the daily patient flows are satisfactory.

2.9.1 Comparison of patient flow metrics considering all patient types between Two-stage SMILP and base-case for case-2 demand scenario

Considering different threshold levels for patient flow metric distributions, Fig 5, 6, and 7 compare direct wait time distributions for providers for case-2 demand scenario between base-case and Two-stage SMILP with threshold levels: λ = 50%, 65%, and 80% quantiles. Following the figures, the average wait time for solutions based on two-stage SMILPs is less than that of base-case, denoting 16%, 6%, and 3% improve for all threshold levels in the SMILPs compared to base-case.

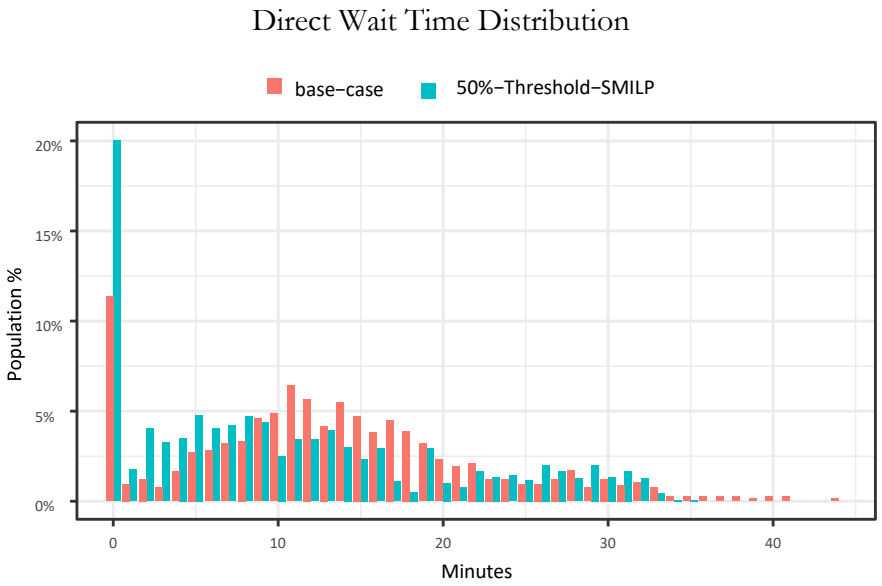

Fig 5. Direct wait time distribution for providers for case-2 demand scenario, comparing base-case and Two-stage SMILP with threshold level: λ = 50% quantile

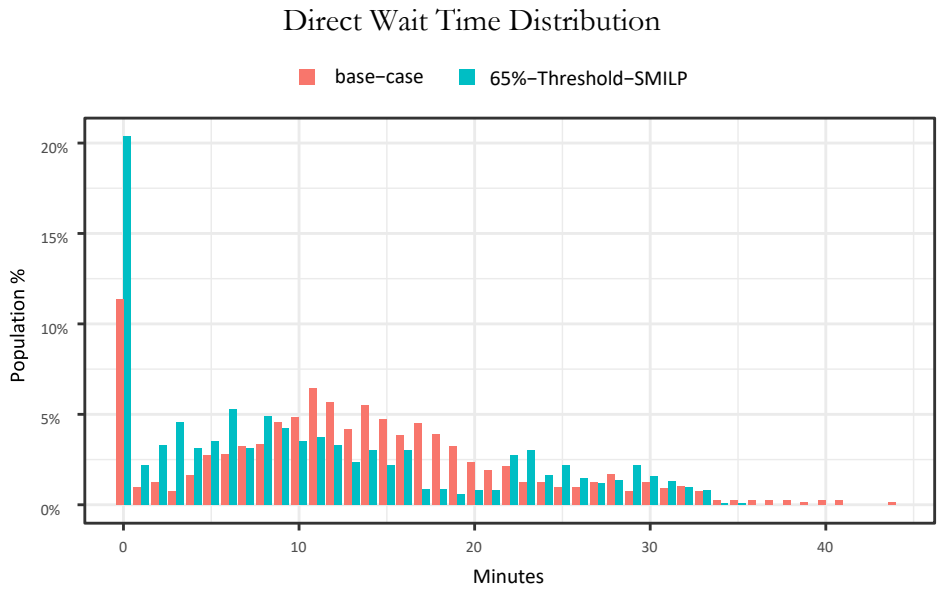

Fig 6. Direct wait time distribution for providers for case-2 demand scenario, comparing base-case and Two-stage SMILP with threshold level: λ = 65% quantile

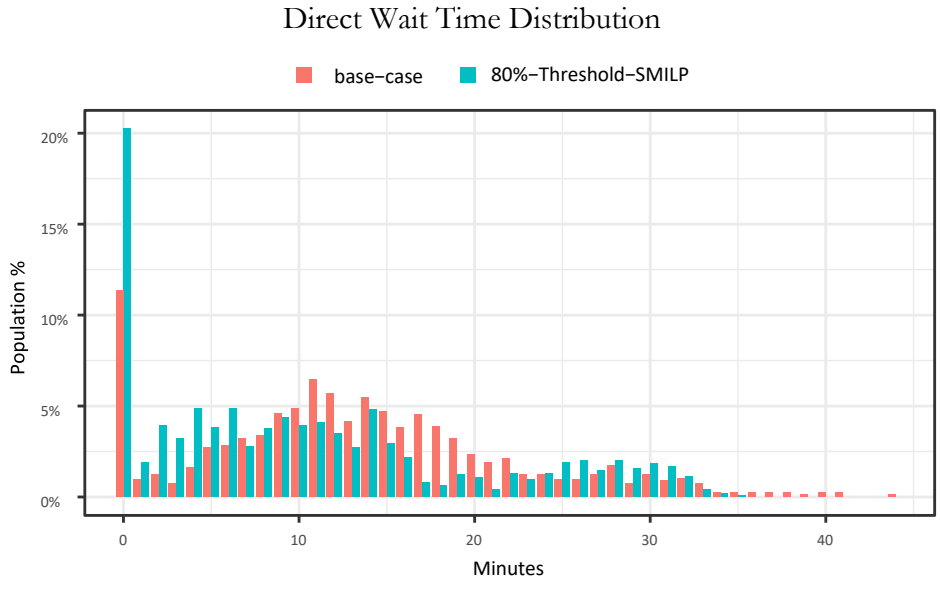

Fig 7. Direct wait time distribution for providers for case-2 demand scenario, comparing base-case and Two-stage SMILP with threshold level: λ = 80% quantile

2.9.2 Comparison of patient flow metrics for patient type-1 between Two-stage SMILP and base-case for case-2 demand scenario

Considering different threshold levels for patient flow metric distributions, Fig 8 compares the direct wait time distributions for patient type-1 for case-2 demand scenario between base-case and two-stage SMILP with threshold levels: λ = 50%, 65%, and 80% quantiles. As shown in the Figure, the average waiting time in two-stage SMILP is less than that of base-case with improving 23%, 19%, and 27% for three threshold levels. At threshold level λ = 50%, median is 20 minutes which is less than other threshold levels 65%, 80% and the base-case in order, which shows that fifty percent of population has a waiting time under 20 minutes. Moreover, 25% of population in threshold level λ = 50% has a waiting time under 1 minute, and between 65% and 80% under 2 minutes, and in comparison to base-case, 25% of the population experience waiting time under 7 minutes which shows two-stage model results are much more robust. The graphs in Fig 8 is for patients with more criticality factor, patient type 1. We refer the graphs for other patient types in appendix.

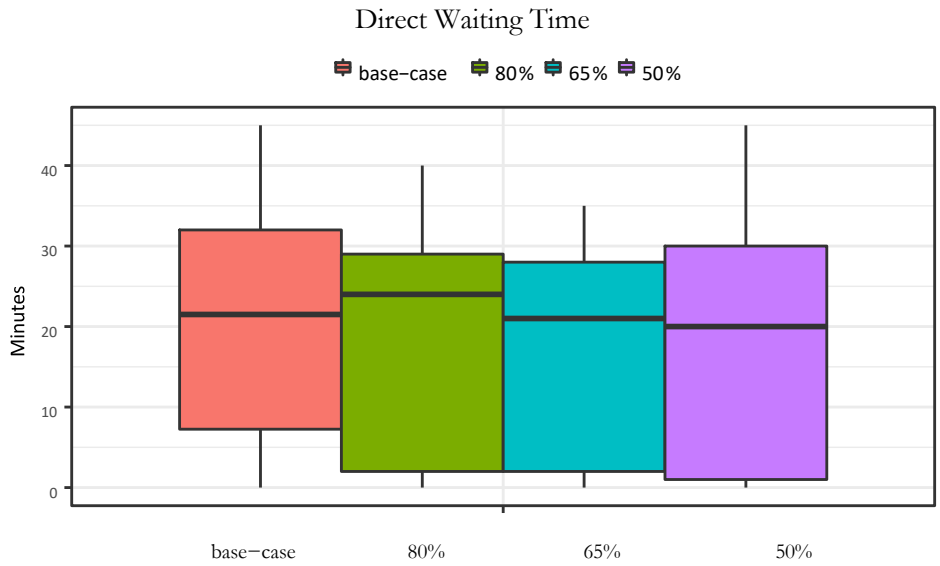

Fig 8. Direct wait time distribution for patient type-1 for case-2 demand scenario, comparing base-case and Two-stage SMILP with threshold level: λ = 50%, 65%, and 80%

2.9.3 Trade-off between direct wait time and indirect wait time considering all patient types between Two-stage SMILP and base-case for case-2 demand scenario

Fig 9 represents the direct wait time distribution for different threshold levels λ = 50%, 65%, and 80% quantiles. Fig 10 compares the average wait time for all threshold levels which shows average wait time is higher in threshold level λ = 80%, 65%, and 50% respectively. In Fig 11 and 12, indirect wait time (delay) distribution, compares two-stage SMILP for case-2 demand scenario with threshold levels: λ = 50%, 65%, and 80% quantiles and base-case. The two-stage SMILP assigns appointment within two weeks while in the base-case, there are assigned appointment for weeks one, two, three, and four. It shows that two-stage SMILP results in less appointment delay compared to the base-case. As per in Fig 10, the average waiting time is higher in the base-case, 80%, 65%, and 50% threshold level in that order. Therefore, we expect the crowded clinic days for indirect wait time in the same order as shown in Fig 11 and 12. Moreover, as represented in Fig 11 for the base-case, we conclude that it has appointment slots with highest waiting times. since it has not only busy clinic days (more appointments in the first 2 weeks) compared to other threshold levels but also it has some appointments for weeks 3 and 4.

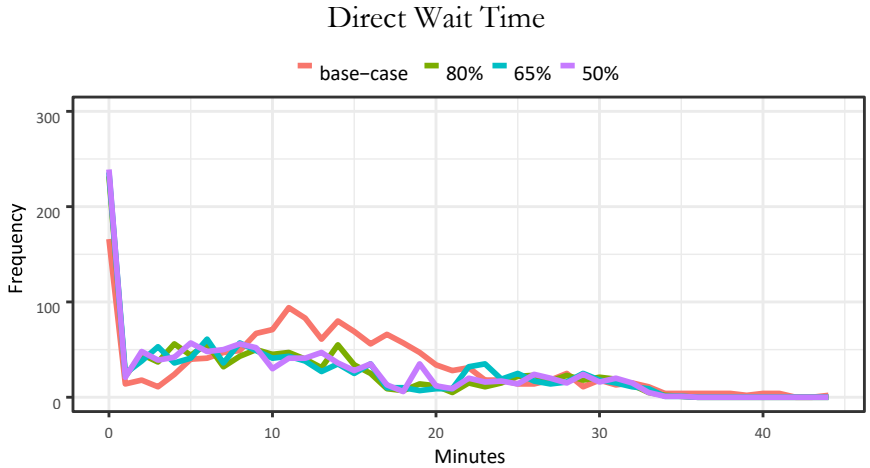

Fig 9. Comparing direct wait time distributions for Two-stage SMILP with threshold levels: λ = 50%, 65%, and 80% quantile and base-case for providers for case-2 demand scenario

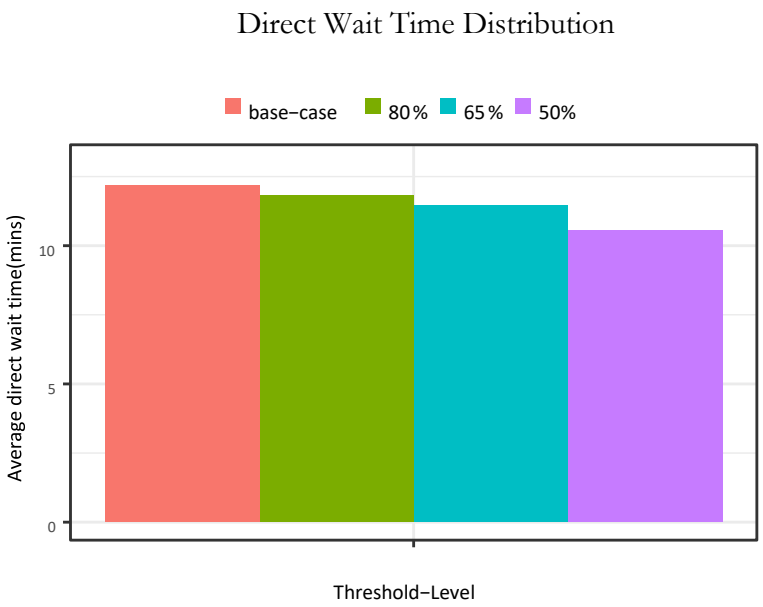

Fig 10. Comparing average wait time for Two-stage SMILP with threshold levels: λ = 50%, 65%, and 80% quantile and base-case for providers under case-2

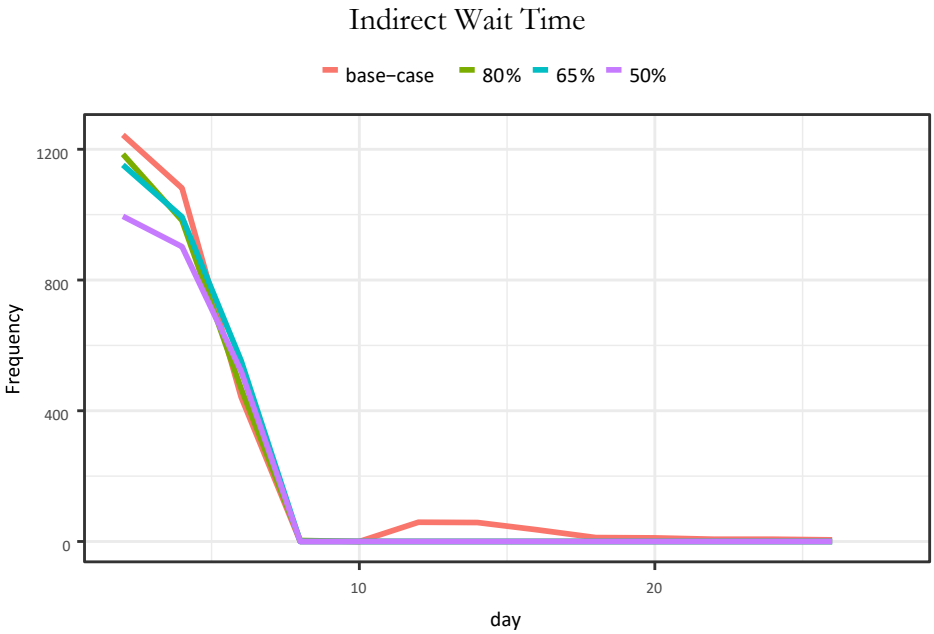

Fig 11. Indirect wait time distribution for case-2 demand scenario, comparing base-case and Two-stage SMILP with threshold levels: λ = 50%, 65%, and 80% quantiles

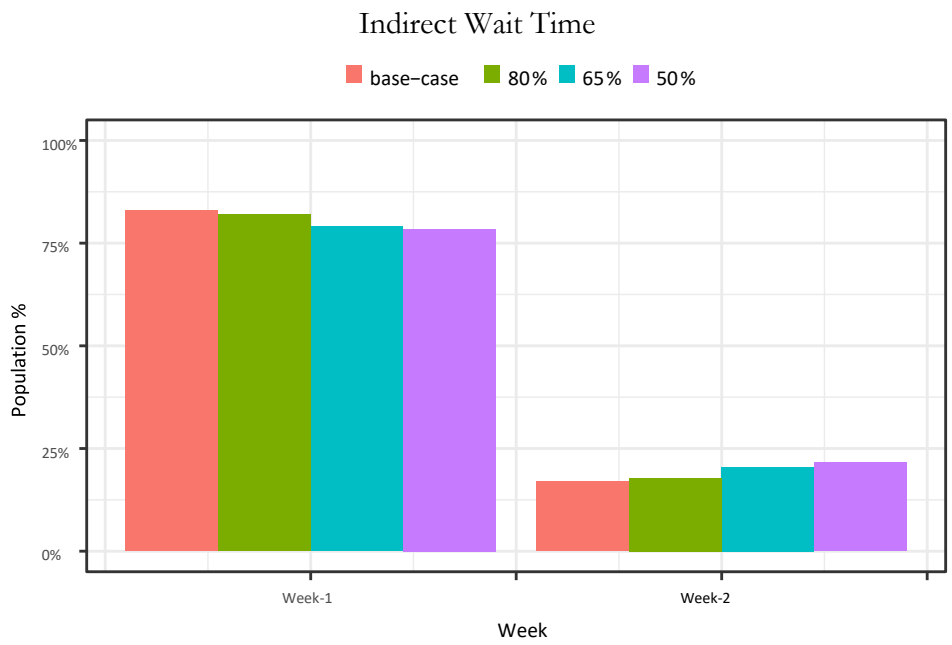

Fig 12. Indirect wait time distribution for case-2 demand scenario, comparing Two-stage SMILP with threshold levels: λ = 50%, 65%, and 80% quantiles and the base-case

2.9.4 Trade-off between direct wait time and indirect wait time for each patient type between Two-stage SMILP and base-case for case-2 demand scenario

Fig 13 represents the trade-off between direct wait time and indirect wait time distributions for SMILP with different threshold levels λ = 50%, 65%, and 80% quantiles and the base-case. The average waiting time is higher in the base-case, 20 minutes, compared to other threshold levels. Therefore, we expect the crowded clinic days for indirect wait time in the beginning of the time horizon compared to other threshold levels. The base-case shows a weaker result compared to the two-stage stochastic programming; since it has more appointments in the first 2 weeks and some appointments for weeks 3 and 4. The graph depicts criticality factors for the patient type 1. We refer the graphs for other patient types in appendix.

Table 6 and 7 show the advantage of using two-stage SMILP over base-case; improving average direct waiting time and indirect waiting time when applying two-stage SMILP.

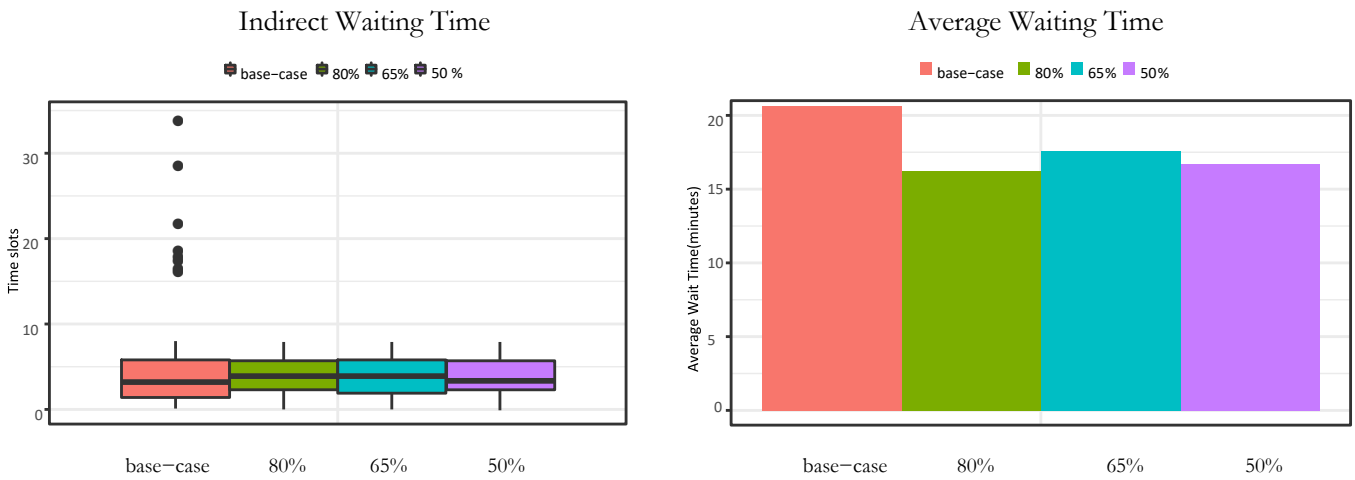

Fig.13. Trade-off between direct wait time and indirect wait time distributions between Two-stage SMILP with threshold levels: λ = 50%, 65%, and 80% quantiles and base-case for patient type-1 for case-2 demand scenario

Threshold Levels	λ = 50%	λ = 65%	λ = 80%
Patient type-1	23%	19%	27%
Patient type-2	11%	10%	10%
Patient type-3	49%	38%	27%
Patient type-4	12%	13%	3%
Patient type-5	20%	31%	13%
Patient type-6	23%	18%	23%
Patient type-7	34%	87%	44%
All patient types	16%	6%	3%

Table 6. Improving average direct waiting time when applying two-stage SMILP

Time-window	λ = 50%	λ = 65%	λ = 80%
Week-1	6%	5%	1.2%
overall	13.5%	12%	7%

Table 7. Improving indirect waiting time when applying two-stage SMILP

2.9.5 Optimal weekly scheduling template

In this section the optimal weekly scheduling template for case-1 and 2 demand scenarios are presented. We calculate and summarize the system's utilization based on the available data in table 4 with available weekly 40 hours for two providers and present them in table 8 below. The statistics shows system reaches steady state in both cases of demand scenarios.

Demand Scenario	Total Weekly Time(mins)	Utilization	Idle time	Avg. appointment time (mins)	Appointment number	Appointment number Our scheduling result
Case-1	1167	48%	52%	11.33	211.82	160
Case-2	2303	96%	4%	11.34	211.55	230

Table 8. Statistics of the system's utilization based on the available data

Table 9 shows the scheduling template of free time slots for providers for office work/ lunch for case-2 demand scenario. This result is the best one in many runs of scenarios. In Table 10 and 11 we calculate the expected service time (minutes) for our scheduling templates using the data in Table 4. Tables 12 and 13 represent the optimal weekly scheduling template for case-1 and case-2 demand scenarios respectively, where NL, FL, FH, NG, MG, EG, and RG stand for New Low-Risk OB, Follow Up Low-Risk OB, Follow Up High-Risk OB, New GYN, MAU GYN, Established GYN, and Results GYN in order.

Slot Index	Mon	Tue	Wed	Thu		Fri	
9							
10							
11							
12							
13							
14							
15							
16							

Table 9. Free time slots for providers for office work/ lunch – case-2 demand scenario

Slot Index	Mon	Tue	Wed	Thu	Fri
1	12	12	12	6	20
2	12	24	12	16	20
3	19	12	24	30	20
4	25	25	16	20	12
5	20	25	12	16	22
6	12	16	27	24	12
7	20	16	16	25	12
8	16	12	21	21	24
9	21	28	16	25	36
10	20	24	20	33	25
11	20	33	24	25	21
12	23	21	16	0	43
13	12	18	25	21	28
14	12	21	16	30	20
15	23	20	24	25	35
16	12	24	12	16	20
Day.average.time(mins)	17.44	20.69	18.31	20.81	23.33

Table 10. Expected service time (minutes) for each time slot – case-1 demand scenario

Slot Index	Mon	Tue	Wed	Thu	Fri
1	18	30	30	35	20
2	22	25	26	38	26
3	26	22	26	24	20
4	26	30	38	20	30
5	29	41	30	20	38
6	26	30	34	35	38
7	30	26	16	35	40
8	26	22	26	44	30
9	29	22	20	23	0
10	16	50	33	45	30
11	22	42	46	24	40
12	23	46	30	20	49
13	33	22	22	34	44
14	22	22	40	20	44
15	26	22	40	35	44
16	38	50	22	34	0
Day.average.time(mins)	25.75	31.375	29.9375	30.375	30.8125

Table 11. Expected service time (minutes) for each time slot – case-2 demand scenario

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

Table 12. Weekly scheduling template for case-1 demand scenario

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

Table 13. Weekly scheduling template for case-2 demand scenario

2.9.6 Sample Average Approximation (SAA)

To estimate a lower bound for risk-neutral SMILP, we choose $\mathcal{N}$ = 20, and 50 scenarios which is repeated M = 20 times. Average of 20 runs is an estimate of lower bound on the objective value. A sample of $\mathcal{N}'$= 1000 scenarios, which is generated independently of the samples were used to get the lower bound, is selected to estimate an upper bound for the optimal solution. In Table 13, upper and lower bounds for objective function value using SAA method is presented. $gap\%$ and $gap_{N,M,N'}$ indicate the differences between upper and lower bounds. Table 14 shows problem with 50 scenarios results in minimum gap percentages which has been used in our experimental settings.

Case	$\mathcal{N}$	Lower bound		Upper bound			
		Average	σ_{LB}	Average	σ_{UB}	$gap_{N,M,N'}$	$gap\%$
CD1	20	191,000	3,235	194,684	2,589	3,684	1.92%
	50	189,000	2,827	192,254	1,925	3,254	1.72%
CD2	20	228,000	2,949	231,152	1,752	3,152	1.38%
	50	228,000	1,883	230,960	1,354	2,960	1.29%

Table 14. Statistical lower and upper bounds of the SAA problems for M = 20 and $\mathcal{N}'$= 1000
CD1: Case-1 Demand, CD2: Case-2 Demand

2.10 Conclusion

In this chapter, we presented methods for improving flow through outpatient clinics focused on OBGYN clinics considering effective appointment scheduling policies by applying Two-Stage Stochastic Mixed-Integer Linear Program Model (two-stage SMILP) approaches to improve patient flow metrics: direct wait time (clinic wait time) and indirect wait time. The mathematical formulation of the problem can be applied to any scheduling modeling in health care that consists of multiple patient types with no-show behavior as well as stochastic servers, follow-up surgery appointments, and overbooking. We model the scheduling problem with many scenarios under certain realization in the second-stage of the problem and examine the effect of this modeling on the first-stage decisions.

Due to the size of the problem instances, a sample average approximation method is used to solve our problem. As we look at the results, we see two-stage SMILP with threshold levels λ = 50%, 65%, and 80% results in better direct and indirect waiting time comparing to the base-case, where average waiting time improved by 16%, 6%, and 3% and indirect waiting time improved by 13.5%, 12%, and 7% for all threshold levels. In our case we run two-stage SMILP once and then use the weekly scheduling template as a guideline for the whole time horizon.

One contribution to this chapter is considering how often the two-stage SMILP needs to be run depending on the available data for seasonality purposes in the different clinics. Another contribution could be on appointment policies in call-center. One may modify the heuristic policy and discuss on different rules in appointment assignment considering multiple patient types along with each type preferences. Next contribution is related to risk-averse models. Risk-averse objectives can be used instead of risk-neutral objectives in order to control the variability of the target performance measures. A few optimization studies propose risk-averse objectives, such as the Markowitz mean-variance method (e.g., (Mak et al. 2015); (Qu et al. 2012)) and the Von Neumann–Morgenstern expected utility method (e.g., (Kemper et al. 2014); (Kuiper & Mandjes 2015); (LaGanga & Lawrence 2012); (Vink et al. 2015)). In the proposed risk-neutral two-stage stochastic model we consider expected value as a performance measure while in a research extension one can use Conditional-Value-at-Risk (CVaR) as a performance measure adding the presence of the risk to the model and evaluate the result.

CHAPTER 3: RISK-AVERSE TWO-STAGE STOCHASTIC PROGRAMMING MODEL TO OPTIMIZE THE PATIENT FLOW METRICS AT OUTPATIENT CLINICS

3.1 Introduction

Risk-neutral two-stage stochastic programming is a long-established approach that has been used in many studies. This method considers the expected value in the objective function as the preference criterion. Moreover, as we discussed in chapter 2, in two-stage stochastic programming the objective function of the second-stage problem, known as the recourse (cost) function, is a random variable. Therefore, the total cost function is a random variable, and determining the optimal decision of the first-stage leads to the problem of comparing random cost variables. However, comparing random variables is one of the main streams in decision theory in the presence of uncertainty in the system, so it is important to consider the effect of variability and specify the preference relations among the random variables using risk measures.

(Daniels et al. 1995) indicate that a critical disadvantage of using the expected value as a performance measure is that it does not account for the risk averse attitude of a decision-maker. In recent years, one of the main approaches in the practice of decision making under risk is mean-risk models, and many researchers have used several varieties of risk measures in their models. (Markowitz 1952) and (De et al. 1992) used variance as the risk measure. The solution to these problems results are inferior, and in the case of a scenario-based approach, the sample variance of any given performance measure involves a quadratic expression, which makes the optimization problem comparatively hard to solve. To remedy this drawback, risk averse approaches are introduced and CVaR is one such approach. CVaR has attracted much attention in recent years. It has been used in research areas such as financial risk management, machine scheduling problems and healthcare (Morgan 1994), (Duffie & Pan 1997), (Ogryczak & Ruszczyński 2002), (Sarin et al. 2014) and (Qi, J., 2017). CVaR simultaneously reduces both the expected value and variance of a performance measure while keeping the linearity whenever the expectation can be represented by a linear expression as in our case. Reported by (Sarin

et al. 2014), it adds benefit over traditional nonlinear problems using expectation-variance methods as well.

In another study by (Schultz & Tiedemann 2003) excess probability has been studied as another risk measure. They confirm that excess probabilities lead to a risk measure which is consistent with the first-degree stochastic dominance relation. They consider linear two-stage stochastic programs with a mixed-integer recourse and propose a scenario decomposition algorithm for computational results.

In this chapter, we investigate the effects of varability in the system by introducing the Conditional-Value-at-Risk (CVaR) as the risk measure and compare the results with expected value approach. In other words, we consider a risk-averse two-stage stochastic programming model, where we specify the Conditional-Value-at-Risk (CVaR) as the risk measure. We believe that this criterion is an effective method to find risk-averse solutions for stochastic programming with applications in scheduling. We apply the proposed model to healthcare operational management, which is one of the research fields that can significantly benefit from risk-averse two-stage stochastic programming models in the presence of uncertainty in demand. We present numerical results to discuss how incorporating a risk measure affects the optimal solutions and demonstrate the computational effectiveness of the proposed methods.

In particular, we consider the problem of determining methods for improving patient flow metrics in outpatient clinics introducing effective appointment scheduling policies by applying the mean-risk Two-Stage Stochastic Mixed-Integer Linear Program (two-stage SMILP) approach is utilized to improve patient flow metrics: direct wait time (clinic wait time) and indirect wait time considering patient's no-show behavior, stochastic server, follow-up surgery appointments, and overbooking. We develop two models: first, a method to optimize the (weekly) scheduling pattern for individual providers that would be updated at regular intervals based on the type and mix of services

rendered, and second, a method for dynamically scheduling patients using the weekly scheduling template. Scheduling will entertain the possibility of arranging multiple appointments at once. The aim is to increase throughput per session while providing timely care, continuity of care, and overall patient satisfaction as well as equity of resource utilization. In chapter two, we developed the risk-neutral approach by minimizing the expected value as a performance criterion without considering risk in the system. However, considering a risk in the model (in the presence of random variables, cost function) is an important factor in healthcare engineering. In chapter three, we model risk-averse two-stage stochastic programming by considering CVaR as a risk measure. For computational results we find the distributions for patient flow metrics and show the advantages of cosidering risk measure in the model.

This chapter is organized as follows. Section 3.2 reviews the relevant literature. Section 3.3 describes model assumptions and framework. Section 3.4 formulates a two- stage mean-risk stochastic programming. Solution of the two-stage SMILP provides the optimal capacity assigned for each time slot. Section 3.5 explains a demand generation simulation. In section 3.6, we introduce a dynamic appointment scheduling policy for actual appointment assignment for different patient types. Section 3.7 explains clinic simulations and direct wait time**.** In this section we calculate the direct wait time experienced by individual providers. Section 3.8 describes the case study and data driven from literature. Section 3.9 provides future research in appointment scheduling.

3.2 Literature review

We categorize the literature review into two sections: first, we briefly review the literature in appointment scheduling focused on outpatient clinics. For a comprehensive review we refer the reader to chapter 2. Then, we investigate surveys using risk measures in objective functions and their advantages over the traditional case using expected value.

Referring to (Ahmadi-Javid et al. 2017), decision making in outpatient appointment scheduling

can be classified into three categories: strategic, tactical, and operational decisions which are long, medium and short-term decisions, in that order. Strategic decisions deal with areas of research on access policy, the number of servers, policy on acceptance of walk-ins, and type of scheduling. On the other hand, tactical/planning decisions determine how several groups of patients are scheduled, and decisions on allocation of capacity to patient groups, appointment slot (interval), appointment scheduling window, and priority of patient groups are made whereas operational decisions are related to scheduling each patient upon his/her request. In other words, it includes decisions related to the allocation of patients to servers/physicians, appointment day/time, patient acceptance/rejection, and patient sequence. The majority of researchers have focused on operational decisions and tactical decisions, but few are available on strategic decisions, which is a broad area for future work.

In general, the performance measure of every health care system involves two aspects: patients' perspectives and providers' perspectives. We aim to improve the performance measure of an outpatient clinic through appointment scheduling considering several criteria: one belongs to patient satisfaction measurement such as waiting time (direct and indirect) which is the most common issue in outpatient appointment scheduling. One commonly used service quality measure for describing this preference is patient expectation. However, the expected waiting time criterion may not satisfactorily distinguish patients' attitudes toward uncertain delays because it corresponds to the average delay experienced by the patient over a potentially infinite number of visits under the same identical conditions. Patient waiting time (direct/indirect waiting time), continuity of care and patient preferences are factors used to measure patient satisfaction. Moreover, patient waiting time, provider over time and provider idle time are the most common performance metrics used in optimization studies. On the other hand, considering indirect waiting time (i.e., the time between the appointment request and the scheduled appointment time) in the objective functions as well as patient preferences in appointment scheduling are mostly referred to future studies as it requires complexity in calculations

(Gupta & Denton 2008), and only a few articles such as (Zacharias & Armony 2016) have this contribution in their work.

(Cartwright et al. 1992) and (McCarthy et al. 2000) declared that a certain waiting time can be acceptable among patients from the patients' perspective and (Camacho et al. n.d.) stated that dissatisfaction with the waiting process may not increase proportionally with the length of the waiting time. In a survey by (Hill & Joonas 2005), 86% of patients accept 30 minutes or fewer as an acceptable threshold for waiting time while in research conducted by (Huang 1994), empirical results disclose patients' acceptable threshold level of waiting time as on average pf 37 minutes, and their patience may decline when the service delay exceeds this threshold. Another perspective is physicians' tolerance; their key performance indicator lies in the proportion of patients seen within a certain time window/threshold level, instead of the total expected waiting time. Reported by the United States and United Kingdom (National Health Service) and (RE, H. 2006), 30 minutes is considered as an acceptable threshold level from patients' perspectives.

In another study by (Toh & Sern 2011) on orthodontic specialist clinics, for those patients arrive on time at the clinic, the percentage of patients that can be seen within 30 minutes of the appointment time should be greater than 50%, whereas in an operating theater of a local hospital in Singapore, less than 30% of patients assigned for surgery experienced more than 30 minutes waiting time. Following these empirical results, some researchers use a tolerance threshold to describe patient satisfaction with waiting processes and take the frequency of delays above this threshold as a service quality measure.

(Qi, J., 2017) proposed a method to address the displeasure of both patients and physicians by balancing the service levels and time measures in the system. A threshold-based performance measure known as Delay Unpleasantness Measure is introduced to assess uncertain delays. Applying this method, the frequency and intensity of a system's satisfaction measures such as patient waiting time

and provider over time is controlled when it is above fixed patient and physician thresholds. As the model considers the threshold for the physician's over time, idle time is being controlled indirectly which is not discussed in her paper. Then, the concept of lexicographic min-max fairness is applied to improve fairness in the appointment scheduling design. In this research information about patients is known prior to the start of the clinic session, which belongs to static appointment scheduling.

Now, we review the surveys using risk measures in the objective values and their advantages over the traditional risk-neutral stochastic programming using the expected value. Integrating risk measures into the objective functions in two-stage stochastic programming is quite recent research. This idea has been used in many studies such as (Ahmed 2004), (Ahmed 2006), (Schultz & Tiedemann 2006), (Fábián 2008), and (Sarin et al. 2014).

For a recent survey on mean-risk stochastic programs, we refer the interested reader to the work of (Krokhmal et al. 2011). In this survey one can review a comprehensive literature review in decision making under uncertainty with the focus on the methods for modeling and controlling of risk in the system.

Using CVaR in model formulation in stochastic scheduling problems which have pervasive applications is an effective approach. As stated by many researchers, (Krokhmal et al. 2011), (Sarin et al. 2014), and (Qi 2017), it will not only reduce both expectation and variance of a performance measure but also when the expectation can be rendered by a linear formulation, it maintains linearity, and this later property has a great advantage over traditional nonlinear expectation-variance-based methods. (De et al. 1992) used variance as a risk measure to determine expectation-variance based efficient schedules. However, using variance as a risk measure has several drawbacks. First, except for some special cases (such as the single machine flow time problem discussed by (De et al. 1992)), it is difficult to derive analytical expressions for the variance of typical performance measures. Moreover, if a scenario-based approach is adopted, the sample variance of any given performance measure

involves a quadratic expression, which makes the optimization problem relatively hard to solve. Second, minimizing the variance of a random variable equally penalizes positive and negative deviations from its mean value.

In research by (Sarin et al. 2014) CVaR is used as a criterion for stochastic programming with applications in scheduling problems. In this paper, a scenario-based MIP model is developed considers CVaR as a risk measure. Then, the method is applied to a single machine as well as in the context of a parallel machine total weighted tardiness problem, and an L-shaped algorithm and a dynamic programming-based heuristic procedure is presented as a solution strategy.

(Ahmed 2004) and (Ahmed 2006) scrutinize different mean-risk objective functions and corresponding computational suitability in addressing risk in stochastic programming models. In these papers Ahmed shows the complexity of mean-variance stochastic programming which leads to NP-hard optimization problems, which is computationally intractable even in the simplest stochastic programs. Next, he introduces several mean-risk functions: the mean-Conditional-Value-at-Risk (CVaR) objective, the mean-semideviation objective, the mean-quantile deviation, and the mean-Gini mean difference objective, which all preserve convexity and are computationally tractable using negligible variants of existing stochastic programming decomposition algorithms. (Schultz & Tiedemann 2006) deals with two-stage mixed-integer stochastic programming and consider Conditional Value-at-Risk as a risk measure. Their model formulation involves the integer variables in the second-stage problem which makes the problem non-convex. Hence, straightforward decomposition algorithms cannot be applied. As a result, they develop the split-variable formulation and a solution algorithm applying the Lagrangian relaxation of non-anticipativity.

In this chapter, there are two levels of decisions: in the first decision, which is advance scheduling, we decide on how many patients to assign within a fixed time slot length, whereas in the second decision, the appointment allocation for each patient is assigned to each time slot. In this research we

consider indirect waiting time as part of the model formulation in the objective function while we consider providers' workloads, and two levels of tactical and operational decisions as integrated models are studied. It means decisions are dependent on each other and are taken simultaneously. Moreover, continuity of care and considering patient preferences as well as direct and indirect waiting time are the flow metrics we measure to evaluate patient satisfaction in our optimization problem, which is unique in terms of methodology. Moreover, while improving metrics on average, we include the CVaR in the model formulation to ensure no subset of patients are experiencing extreme waiting times and compare the results with a case in its absence. In other words, we compare the results of risk-neutral two-stage stochastic programming and risk-averse two-stage stochastic programming to present the advantages of using CVaR. Using CVaR in model formulation in stochastic scheduling problems has benefits; it will reduce both expectation and variance of a performance measure and at the same time retains linearity whenever the expectation can be presented by a linear expression. We begin by formulating a scenario-based (stochastic) mixed integer linear programming to minimize CVaR for outpatient appointment scheduling. For the solution scheme we use sample average approximation (SAA) to decide on the number of scenarios needed for our calculations. Next, we calculate the performance measure: direct and indirect waiting time. In the first phase of our research we find the optimal weekly scheduling template as a result of our tactical decisions, and in the second phase, we make the operational decisions by dynamically assigning an appointment to an arriving patient's requested time.

3.3 Model assumptions and framework

In this chapter we design an appointment scheduling model that channels multiple patient types to a team of providers in a women's specialty clinic. The objective is to improve patient flow through outpatient clinics using efficient appointment scheduling policies. Recent research suggests that continuity of care not only results in patient satisfaction but also improves the patient health

specially when patient's health condition is in the early stages. In order to reach this goal, we aim to minimize the indirect waiting time in the model formulation as part of the objective function and direct waiting time at the clinic specialty as part of our constraints in our model. Direct waiting time, known as clinic delay, is physical waiting time experienced by the patients once they arrive at the clinic; indirect waiting time, known as appointment delay, is defined as the time window between the appointment request and the offered appointment, (Zacharias & Armony 2016).

The process for decision making includes three steps: in the first step, which we call the tactical decision, we optimize the maximum capacity for the scheduling template which entertains channeling multiple patient types to the provider team. The objective in the first-stage is to balance the provider workload between day sessions as well as among each time slot. In the second step, we create the operational model, a dynamic appointment scheduling which assigns appointment time to a patient request. In the third step, we evaluate the appointment system by a feedback decision; we check the daily average waiting time of the sequence of patients and if it is higher than the accepted threshold level, we remove that sequence from the tactical decision (Fig 14). We refer the reader to chapter two of this dissertation for more information on the patient types and process.

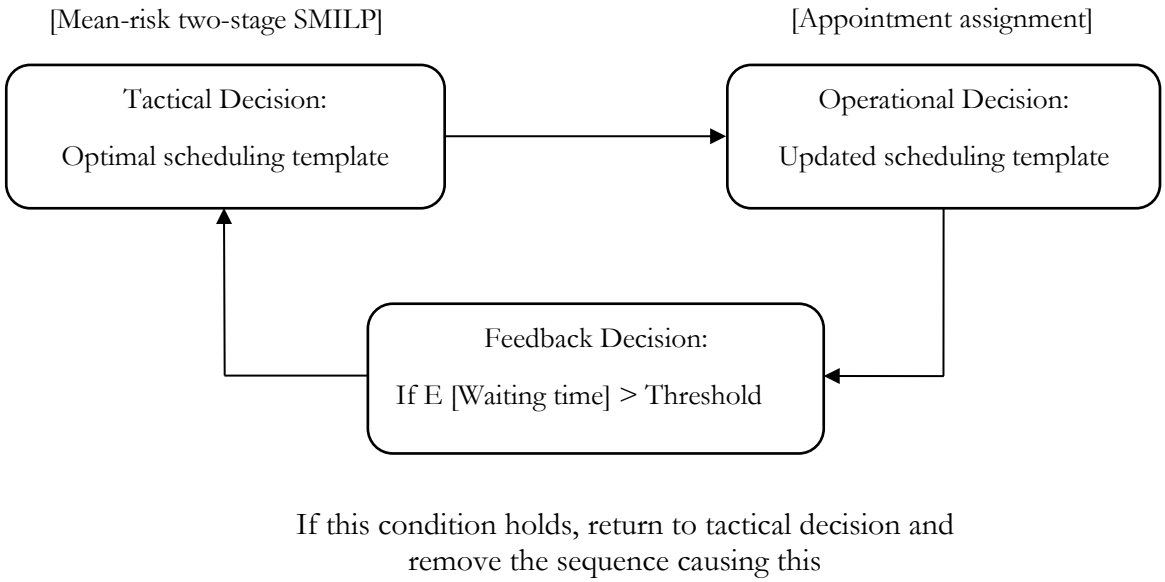

Fig 14. Research framework

Some clinics group two to four physicians as a provider team to improve continuity of care with scheduling flexibility. Referring to (Qu et al. 2013), we assume a team of two providers. Patients are scheduled with any available provider in each clinic session (morning/afternoon) with identical service slots of 15 minutes, which is common in practice. Moreover, as many providers are different in their practice styles in specialty clinics, the model considers free capacity for lunch hours/ office work for the provider team and in some cases appointments for follow-up surgery. Service time duration for each patient type is derived from (Qu et al. 2013) and (Lenin et al. 2015). The research framework is the same as shown in Figure 3 in Chapter 2. The contribution of the model goes to the mathematical formulation which we will present in section 3.4.

3.4 Two- stage mean-risk stochastic programming

Stochastic models, which have considered expectation in the objective function make the model formulation risk-neutral. As discussed in the literature review section, to consider the effect of risk in the model outcomes, a risk measure is added to the risk-neutral objective function which is called the mean-risk stochastic program. We use CVaR as a risk measure since minimizing CVaR in two-stage stochastic programming maintains linearity and results in a convex optimization problem that allows to use the easily available convex optimization methods. As an application of this risk measure in financial risk management, suppose X shows the value of a financial position (such as assets, liabilities and owners' equity as at a specific date), its Value-at-Risk at a 0.05 confidence level, denoted as $VaR_{0.05}(X)$, defines the risk of X as the amount that can be lost with probability of no more than 5%, over the given time horizon (e.g., weekly/monthly). In this section we briefly review the risk-neutral two-stage stochastic linear programming and next introduce the model formulation of a two-stage mean-risk stochastic programming framework. For the following definitions and terminology we refer to (Noyan 2012) and (Krokhmal et al. 2011).

Suppose $(\Omega, \mathcal{F}, \mathrm{P})$ is a probability space, where Ω is the sample space, $\mathcal{F}$ is a σ-algebra on Ω

and P is a probability measure on Ω. We consider a finite probability space, where $\Omega = \{\omega^1, \dots, \omega^N\}$ with corresponding probabilities $p^1, \dots, p^N$. The general form of the risk-neutral two-stage stochastic linear programming problem is presented by

$$min_{x\epsilon X}\mathbb{E}[f(\boldsymbol{x},\omega)] = min_{x\epsilon X} c^T\boldsymbol{x} + \mathbb{E}[Q(\boldsymbol{x},\xi(\omega))], \quad (15)$$

where $f(\boldsymbol{x},\omega) = c^T\boldsymbol{x} + Q(\boldsymbol{x},\xi(\omega))$ is the total cost function of the first-stage problem and

$$Q(\boldsymbol{x},\xi^s) = min_{y^s}\{(q^s)^T y^s : T^s\boldsymbol{x} + W^s y^s = h^s, y^s \geq 0\} \quad (16)$$

is the second-stage problem corresponding to the realization of the random data $\xi(\omega)$ for the elementary event ω^s, represented by $\xi^s = (q^s, T^s, W^s, h^s)$. In (2) $\boldsymbol{x}$ and y denote the vector of the first-stage and second-stage decision variables, in that order. We assume all the matrices meet the suitable dimensions and equations (15), (16) and the objective functions are linear. $X \subset \mathbb{R}^n_+$ is a non-empty set of feasible decisions, $Q(\boldsymbol{x},\xi(\omega)) > -\infty$ for all $\omega\epsilon\Omega$, and the second-stage problem (16) maybe infeasible for some first-stage decision $\boldsymbol{x}\epsilon X$. Observe that the first-stage decisions are deterministic, and the second-stage decisions are allowed to depend on the elementary events, i.e., $y^s = y(\omega^s), s = 1, \dots, N$. Basically, the second-stage decisions denote the operational decisions and change depending on certain realizations of the random data. The objective function $Q(\boldsymbol{x},\xi(\omega))$ of the second-stage problem (16), known as the recourse (cost) function, is a random variable; therefore, the total cost function $f(\boldsymbol{x},\omega)$ is a random variable. In conclusion, the optimal decision variable $\boldsymbol{x}$ results in a problem of comparing random cost variables $\{f(\boldsymbol{x},\omega)\}_{x\epsilon X}$ which is one of the main streams of decision theory under uncertainty, and it is essential to consider the effect of variability and add risk measures to the model. One of the important methods in decision making considering risk uses mean-risk models. In these models the minimization is over the mean-risk objective function with a risk measure.

The risk-averse model is represented as the following minimization:

$$min_{x\in X}\{\mathbb{E}[f(\boldsymbol{x},\omega)] + \lambda\rho(f(\boldsymbol{x},\omega))\}$$

while $\rho: Z \to \mathbb{R}$ is defined as the risk measure, where ρ is a function and Z is a linear space of $\mathcal{F}$-measurable functions on the probability space $(\Omega, \mathcal{F}, \mathrm{P})$; λ is a non-negative trade-off coefficient denoting the exchange rate of the mean cost for the risk/weight factor that quantifies the tradeoff between the expected cost and risk, which is also known as the risk coefficient and is determined by decision makers according to their risk preferences. There are many downside risk measures; we refer the readers to (Ahmed 2006) for the complete list. However, we use the Conditional-Value-at-Risk (CVaR) in our model as we explained in the introduction and literature review sections in terms of its application and benefits.

We state that the decision variable $\boldsymbol{x}$ is efficient in the concept of the mean-risk if and only if for a given level of expected cost, $f(\boldsymbol{x},\omega)$ has the lowest possible CVaR, and for a given level of CVaR it has the lowest possible expected cost. One can construct the mean-risk efficient frontier by finding the efficient solutions for different risk coefficients. Thus, we consider the following two-stage mean-risk stochastic programming problem:

$$min_{x\in X}\{\mathbb{E}[f(\boldsymbol{x},\omega)] + \lambda\, CVaR_{\alpha}(f(\boldsymbol{x},\omega))\}, \quad (17)$$

where $CVaR_{\alpha}$ represent the conditional-value-at-risk at level α.

Definition 1. Let $F_z(.)$ denote the cumulative distribution function of a random variable Z.

Referring to the financial literature, the α -quantile

$$\inf\{\eta\epsilon\mathbb{R} : F_z(\eta) \geq \alpha\}$$

is called the value-at-risk (VaR) at the confidence level α and represented by $VaR_{\alpha}(Z), \alpha \in (0,1]$.

Definition 2. The conditional value-at-risk which is called mean excess loss or tail VaR, at level α is defined as

$$CVaR_{\alpha}(Z) = \mathbb{E}[Z|Z \geq VaR_{\alpha}(Z)]. \quad (18)$$

This definition provides a clear understanding of the concept of $CVaR$: $CVaR_{\alpha}(Z)$ is the conditional expected value exceeding the value-at-risk at the confidence level α. In the minimization of the cost function, VaR_{α} is the α-quantile of the distribution of the cost, and it provides an upper bound that is exceeded only with a probability of $1-\alpha$. On the other hand, $CVaR_{\alpha}(Z)$ is a measure of severity of the cost if it is more than $VaR_{\alpha}(Z)$.

Definition 3. The conditional-value-at-risk of a random variable Z at the confidence level α is defined by

$$CVaR_{\alpha}(Z) = inf_{\eta \epsilon \mathbb{R}}\{\eta + \frac{1}{1-\alpha}\mathbb{E}[(Z-\eta)_{+}], \quad (19)$$

where we let $(Z)_{+} = \max\{0, Z\}, z \epsilon \mathbb{R}$. It is well-known that the infimum in (19) is obtained at α-quantile of Z.

In the following subsections, we first introduce model formulation in section (3.4.1). Next, in section (3.4.2) we develop solution scheme; sample average approximation (SAA).

3.4.1 Model formulation

In this section we develop the model formulation for two-stage risk-averse stochastic programming. We use the same assumptions and terminology as used in the risk-neutral model in chapter 2 and develop the contribution on the formulation for a risk-averse model as follows.

The objective of the decision-making problem in the first-stage is to balance a provider's workload not only among morning/afternoon sessions, but also in each time-slot of the clinic. In our model formulation, the first-stage determines the amount of maximum capacity reserved for each patient type assigned to each provider for individual time-slots in a weekly pattern which will be used as a guide for the whole time horizon. In the second-stage, we determine the time-slot utilization for an

individual patient type assigned to each provider for individual time-slots under certain realization ω.

We use the notations shown in Table 15 for the model formulation.

Set	
T	Set of planning horizon
$\mathcal{R}$	Set of providers
$\mathcal{N}$	Set of patient types
N'	Set of new patient type
Ω	Set of all scenarios
RPt	Set of risk factors for different patient type
RPr	Set of risk factors for different provider levels
$\mathcal{H}$	Set of free time slots for each provider over time horizon T
$\mathcal{S}$	Set of morning/afternoon sessions over time horizon T
μ	Set of feedback sequence over morning/afternoon session of every day
β	Set of patients scheduled for specific clinic day
ξ	Set of exam rooms in the clinic
Γ	Set of call, desired and appointment times, indexed by $\gamma(t)\epsilon\Gamma$ containing time-slot, $t\epsilon T$
Parameter	
a_j	Number of new patients desired by provider, $j\epsilon\ \mathcal{R}$
cf_i	Risk factor for patient type, $i\epsilon\mathcal{N}$
CF_j	Risk factor for provider, $j\epsilon\ \mathcal{R}$
tlr_j	Tolerance factor of provider, $j\epsilon\ \mathcal{R}$
Δ_j	Cost of additional capacity of provider, $j\epsilon\ \mathcal{R}$
ρ_j	Cost of new patient type for provider, $j\epsilon\ \mathcal{R}$
c_j	Free capacity for provider, $j\epsilon\ \mathcal{R}$ over time horizon T
p_i	Average no-show probability of patient type, $i\epsilon\mathcal{N}$
α_ω	Probability of scenario, $\omega\epsilon\Omega$

M	A large number
$\mathcal{G}$	Number of time-slots per week
$\lvert\mathcal{S}\rvert$	Cardinality of $\mathcal{S}$
λ	Penalty parameter for penalty variable for each time-slot, $t\epsilon T$
λ_c	Risk/trade-off Coefficient
α	Confidence level, (0,1]
$d_{i,j,\gamma(t)}(\omega)$	Demand of patient-type, $i\epsilon\mathcal{N}$ ask for provider, $j\epsilon\ \mathcal{R}$, with call and desired time set $\gamma(t)\epsilon\Gamma$ under scenario, $\omega\epsilon\Omega$
First-stage decision variables	
$x_{i,j,t}$	Number of patient type, $i\epsilon\mathcal{N}$ assigned to provider, $j\epsilon\mathcal{R}$ per time-slot, $t\epsilon T$
e_j	Penalty variable for provider, $j\epsilon\ \mathcal{R}$ w.r.t. new patient type
$z_{j,t}$	1 if time-slot, $t\epsilon T$ is free for provider, $j\epsilon\ \mathcal{R}$, else 0
dev_t	Penalty variable for each time-slot, $t\epsilon T$
η	Value-at-Risk (VaR), α-quantile, Target level
Second-stage decision variables	
$\upsilon(\omega)$	Auxiliary variable for CVaR for $\omega\epsilon\Omega$
$b_{j,t}\ (\omega)$	Capacity slack variable for provider, $j\epsilon\ \mathcal{R}$, time-slot, $t\epsilon T$, under scenario, $\omega\epsilon\Omega$
$y_{i,j,\gamma(t)}\ \ (\omega)$	Time slot utilization for number of type, $i\epsilon\mathcal{N}$ patient asked for provider, $j\epsilon\ \mathcal{R}$ with call, desired and appointment time set $\gamma(t)\epsilon\Gamma$ under scenario, $\omega\epsilon\Omega$

Table. 15. Notation used in Mean-Risk two-stage SMILP model

First-stage objective function:

$$\min \sum_{j\,\epsilon\,\mathcal{R}} \rho_j \ .\ e_j + \lambda\,.\sum_{t\epsilon T} dev_t \ + \ E_{\Omega}[\,f(x,\tilde{\omega})] + \lambda_c.\eta \qquad (P')$$

First-stage constraints:

$$e_j + \sum_{t\epsilon T}\sum_{i\epsilon N'} x_{i,j,t} \geq a_j \qquad N' \subset N, \forall j \in \mathcal{R} \quad (20)$$

$$e_j - \sum_{t\epsilon T}\sum_{i\epsilon N'} x_{i,j,t} \geq -a_j \qquad N' \subset N, \forall j \in \mathcal{R} \quad (21)$$

$$dev_t - \sum_{j\epsilon\mathcal{R}}\sum_{i\epsilon N} x_{i,j,t} + \sum_{i\epsilon N}\sum_{j\epsilon\mathcal{R}}\sum_{t\epsilon T} x_{i,j,t} / \mathcal{g} \geq 0 \qquad \forall t \in T \quad (22)$$

$$dev_t + \sum_{j\epsilon\mathcal{R}}\sum_{i\epsilon N} x_{i,j,t} - \sum_{i\epsilon N}\sum_{j\epsilon\mathcal{R}}\sum_{t\epsilon T} x_{i,j,t} / \mathcal{g} \geq 0 \qquad \forall t \in T \quad (23)$$

$$\sum_{i\epsilon N} cf_i . x_{i,j,t} \leq CF_j \qquad \forall t \in T, j \in \mathcal{R} \quad (24)$$

$$\sum_{t\in\mathcal{S}}\sum_{i\epsilon N} cf_i . x_{i,j,t} \leq |\mathcal{S}|CF_j - tlr_j \qquad \mathcal{S} \subset T, \forall j \in \mathcal{R} \quad (25)$$

$$\sum_{t\epsilon\mathcal{H}} z_{j,t} = c_j \qquad \mathcal{H} \subset T, \forall j \in \mathcal{R} \quad (26)$$

$$\sum_{i\epsilon N} x_{i,j,t} \leq \mathrm{M}.(1 - z_{j,t}) \qquad \forall t \in \mathcal{H} \subset T, j \in \mathcal{R} \quad (27)$$

$$\sum_{i,j,t\epsilon\eta} x_{i,j,t} \leq |\eta| - 1 \qquad \eta \subset \beta, \eta \neq \emptyset \quad (28)$$

$$x_{i,j,t} \in \mathbb{Z}^+, e_j \in \mathbb{R}^+, z_{j,t} \in \{0,1\}, dev_t \in \mathbb{Z}^+, \mathcal{g} \in \mathcal{G} \quad (29)$$

Second-stage objective function:

$$f(x,\omega) = \min \sum_{i\epsilon N}\sum_{j\in\mathcal{R}}\sum_{\gamma(t)\epsilon\Gamma} y_{i,j,\gamma(t)}(\omega).d_{i,j,\gamma(t)}(\omega).\vartheta + \sum_{j\in\mathcal{R}}\sum_{\gamma(t)/\{tc,td\}\epsilon\Gamma} b_{j,\gamma(t)}(\omega).\Delta_j + \frac{\lambda_c}{1-\alpha}.\upsilon(\omega)$$

Second-stage constraints:

$$\sum_{\gamma(t)/\{ta\}\epsilon\Gamma}(1-p_i).y_{i,j,\gamma(t)}(\omega).d_{i,j,\gamma(t)}(\omega) \leq x_{i,j,t} + b_{j,t}(\omega) \qquad \forall i \in N, j \in \mathcal{R}, t \in T \quad (30)$$

$$\sum_{\gamma(t)/\{tc,td\}\epsilon\Gamma} y_{i,j,\gamma(t)}(\omega) = 1 \qquad \forall i \in N, j \in \mathcal{R}, \gamma(t)/\{ta\}\epsilon\Gamma \quad (31)$$

$$-\sum_{j\in\mathcal{R}}\rho_j . e_j - \lambda . \sum_{t\epsilon T} dev_t - \quad (32)$$

$$(\sum_{i\epsilon N}\sum_{j\in\mathcal{R}}\sum_{\gamma(t)\epsilon\Gamma} y_{i,j,\gamma(t)}(\omega).d_{i,j,\gamma(t)}(\omega).\vartheta + \sum_{j\in\mathcal{R}}\sum_{\frac{\gamma(t)}{\{tc,td\}\epsilon\Gamma}} b_{j,\gamma(t)}(\omega).\Delta_j) + \eta + \upsilon(\omega) \geq 0$$

$$0 \leq y_{i,j,\gamma(t)}(\omega) \leq 1, b_{j,t}(\omega) \in \mathbb{R}, \upsilon(\omega) \in \mathbb{R}^+, \forall\, \omega\epsilon\Omega \quad (33)$$

In the above formulation, constraints (20) and (21) check the difference between the desired number of new patients by individual providers and the assigned number of new patients to each provider. In other words, the equity of new patients among all providers is being evaluated by

constraints (20) and (21). Constraints (22) and (23) calculate all capacities reserved for each time-slot and find the average of the capacities reserved over the week. Finally, they find the deviation between capacities reserved for each time-slot and average the amount over the week. Next, this deviation is penalized in the objective function (P'). In constraint (24), provider workload in each time slot of the clinic is controlled, and individual patient type is channeled to each provider. However, constraint (25) is used to balance the provider workload among clinic sessions while channeling patient types to the providers. Constraint (26) opens free capacity for each provider based on the desired number of time slots by individual providers through afternoon sessions. These free capacities are reserved for emergency/ post-surgery follow-up appointment requests. Constraint (27) guarantees there will be no assignments in time slots obtained by constraint (26). Constraint (28), which is called the feedback *constraint*, is to remove the sequence of patients whose clinic wait time threshold has been violated. In the second-stage, constraint (30) doesn't allow each time-slot utilization to exceed the capacity reserved in the first-stage mixed-integer linear problem. In the second-stage, capacities are determined based on first-stage decisions.

Constraint (31) assigns an appointment time to each demand arrival. Constraint (32) preserves the risk-averse properties. Objective function (P') in the two-stage mixed-integer linear problem penalizes the system's over/under utilization in terms of the time slot. In the first part of the objective function, the model penalizes the over/under utilization of time slots reserved for new patient types for an individual provider as well as all time slot capacities, and in the second part of the objective function, indirect waiting time (the time between a patient's desired time and the assigned appointment time) in terms of time slot is penalized. In the second-stage objective function, ϑ denotes $f(ta - tc).(ta - td)$, where $f(ta - tc) = (ta - tc)^{-\frac{1}{2}}$ is called the penalty function and controls the indirect waiting time of the system; tc and ta denote call and appointment times, respectively. By

changing the trade-off coefficient λ_c the efficient appointment schedule and appointment policies can be constructed, and this would allow the decision maker to evaluate different policies.

3.4.2 Solution scheme: Sample Average Approximation (SAA)

The mean-risk two-stage SMILP solvers can typically solve instances with a small number of scenarios. However, a typical problem instance in a practical case would have thousands of scenarios. Using the sample average approximation (SAA) method is a way to handle this problem.

We use the sample average approximation (SAA) to reduce the size of the problem by repeatedly solving it with a smaller set of scenarios. We generate random samples with $\mathcal{N} < |\Omega|$ realizations of the uncertain parameters and approximate the expected recourse costs by the sample average function $\frac{1}{\mathcal{N}} \sum_{n=1}^{\mathcal{N}} f(x, \tilde{\omega})$. For the complete formulation we refer the reader to chapter 2 of this dissertation.

3.5 Demand Generation

Demand is an input parameter in the mean-risk two-stage SMILP model. We assume the number of patients asking for an appointment is uncertain, so we generate demand for appointment requests for many scenarios. Demand is generated with respect to the following scenarios. We assume a six-month time horizon for our demand generation. The domain for patient calls has been considered for the first four months and their desired time has been generated from a patient call time until the end of time horizon (six months). For more explanations we refer to chapter 2 of this dissertation.

3.6 Dynamic Appointment Scheduling

After finding an optimal weekly appointment scheduling pattern from the mean-risk two-stage SMILP model, the call center uses the solution from the mean-risk two-stage SMILP on daily dynamic appointment assignment. This is referred to as *Call Center appointment assignment.* Next, we simulate the call center with demand generation and develop a heuristic policy to assign an appointment time for each patient's arrival. Patients are quoted their appointment times when they request an appointment.

The sequence of appointments may change over time as the appointment schedule evolves; however, we assume that once an appointment time is assigned for a given patient, it cannot be changed. Our demand generation has these parameters: patient type, provider, call time, and desired time for one scenario. We design *Index heuristic policy* to assign an appointment as follows. We divide the appointment policy into three categories: first week, one month, and a remaining time window. When a patient requests an appointment, it is offered with respect to the maximum capacity in the first week of patient's desired time. If the appointment is not accepted by the patient within the first week, the next appointment time is offered at the earliest availability respective to the patient's request in the remaining month; then, if patient still doesn't accept, we offer the next available time slot in the remaining time window until the patient accepts the appointment time. We summarize the index heuristic policy below.

Index heuristic policy:

Input weekly appointment scheduling template *S*, demand set *D* for time horizon T,
and appointment acceptance threshold τ

1: **for** demand arrival *D* in day i:
2: $Capacity \leftarrow \{\}$
3: **for** $t \in \{DT, \dots, DT+T\}$:
4: find the corresponding capacity for time slot t, $I_t = x_t$, $Capacity = I_t$
where DT is patient's desired time, T is one-week time window,
5: **for** $j \in length\{Capacity\}$:
6: find $t^* = argmax(I_t), t \in \{DT, \dots, DT+T\}$and offer time slot t^* to the patient,
7: If τ meets, update S: $I_t = x_t - 1$ and go to step 1;
otherwise, go to step 8
8: **for** $t \in \{DT, \dots, DT+T'\}$:
9: search the first available slot, $I_t = x_t, x_t > 0, t \in \{DT, \dots, DT+T'\}$, where T'
is one-month time window
10: If τ meets, update S: $I_t = x_t - 1$ and go to step 1;
otherwise go to step 11;
11: **for** $t \in \{DT, \dots, DT+T''\}$:
12: assign appointment slot in the remaining time horizon T'', for $I_t = x_t, x_t > 0$, update
S: $I_t = x_t - 1$ and go to step 1.

Output: updated weekly appointment scheduling template S.

Note that the remaining time window threshold after the first week horizon depends on the patient's urgency. For some patients we may need to consider one month, whereas for other patient types this threshold could be in months. It depends on the patient service category.

3.7 Clinic simulation

As we discussed in the literature review section, most of the research done on outpatient clinics aims to minimize the direct waiting time of the clinic in the model formulation of two-stage mixed integer programming. However, we monitor the clinic waiting time of the system by simulating the clinic using the following formulation. We check the daily expected waiting time of the clinic for a sequence of patients for a given day. After each day, we check if the expected waiting time of the clinic for the given day is greater than some threshold; we avoid creating such a sequence of patients in the future of the planning horizon by removing that sequence. This approach will affect other flow metrics such as the system's over time and idle time. The clinic has multiple servers, and service times in each server are random variables. Figure 15 depicts the resources at every stage of an outpatient procedure clinic.

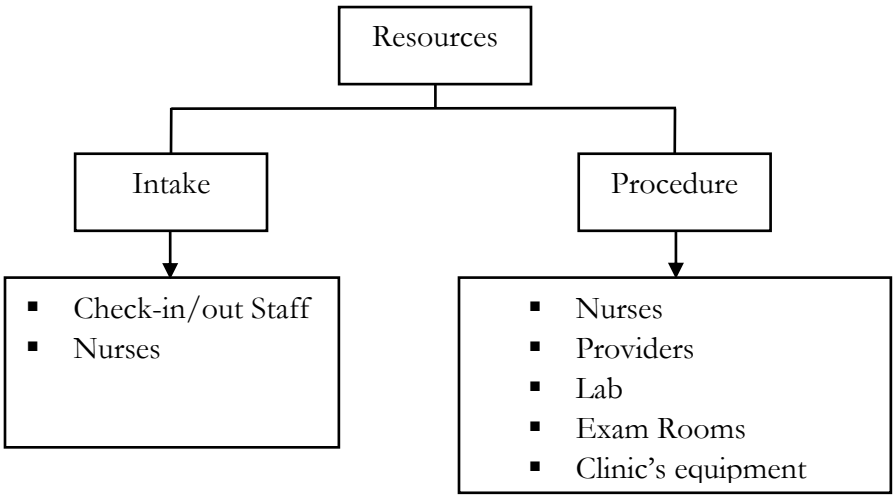

Fig 15. The resources at every stage of an outpatient procedure clinic

We calculate patient waiting time $W_{i,k}$ by developing the following formula considering multiple servers in the system:

$$W_{1,k} = 0 \quad , \forall\, k = 1, \dots, k$$
$$W_{i,k} = (W_{i-1,k} + Z_{i-1,k} - x_{i-1,k})^{+} \quad , i = 2, \dots, n, k = 1, \dots, k$$

where $Z_{i,k}$ is the independent and identically distributed service duration for patient i at clinic room k, $x_{i,k}$ is customer allowance (inter-arrival time between patient i and $i+1$), $(.)^{+}$ indicates $\max(.,0)$ and d is session length. The total waiting time of the system for a given day equals $\sum_{i \epsilon \beta} \sum_{k \epsilon \xi} W_{i,k}$, where β is the set of patients scheduled for an individual clinic day, ξ is the set of clinic rooms in the clinic, k is the number of clinic rooms, and n is the number of patients. The flow of patients at the clinic is shown in Fig 16 and 17. Fig 19 shows an example of a clinic layout at an OBGYN clinic.

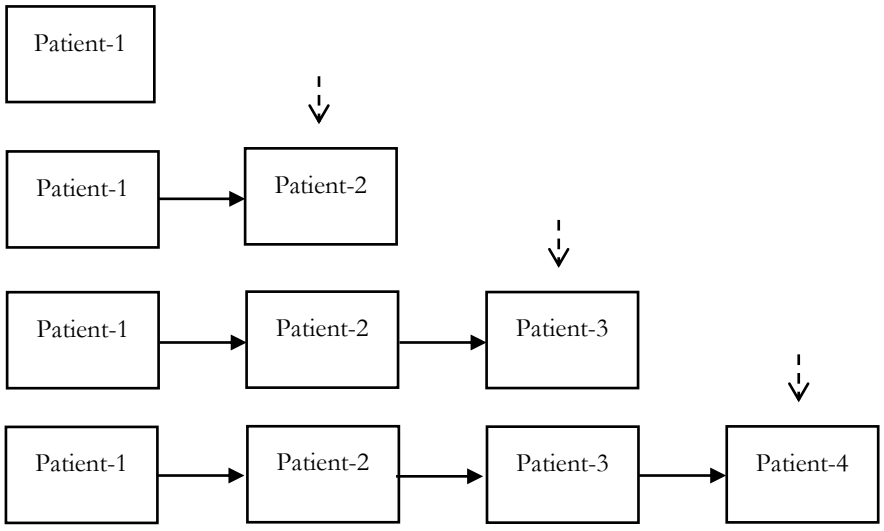

Fig 16. Appointment services in which the sequence of appointments is FCFS (First Come-First Serve)

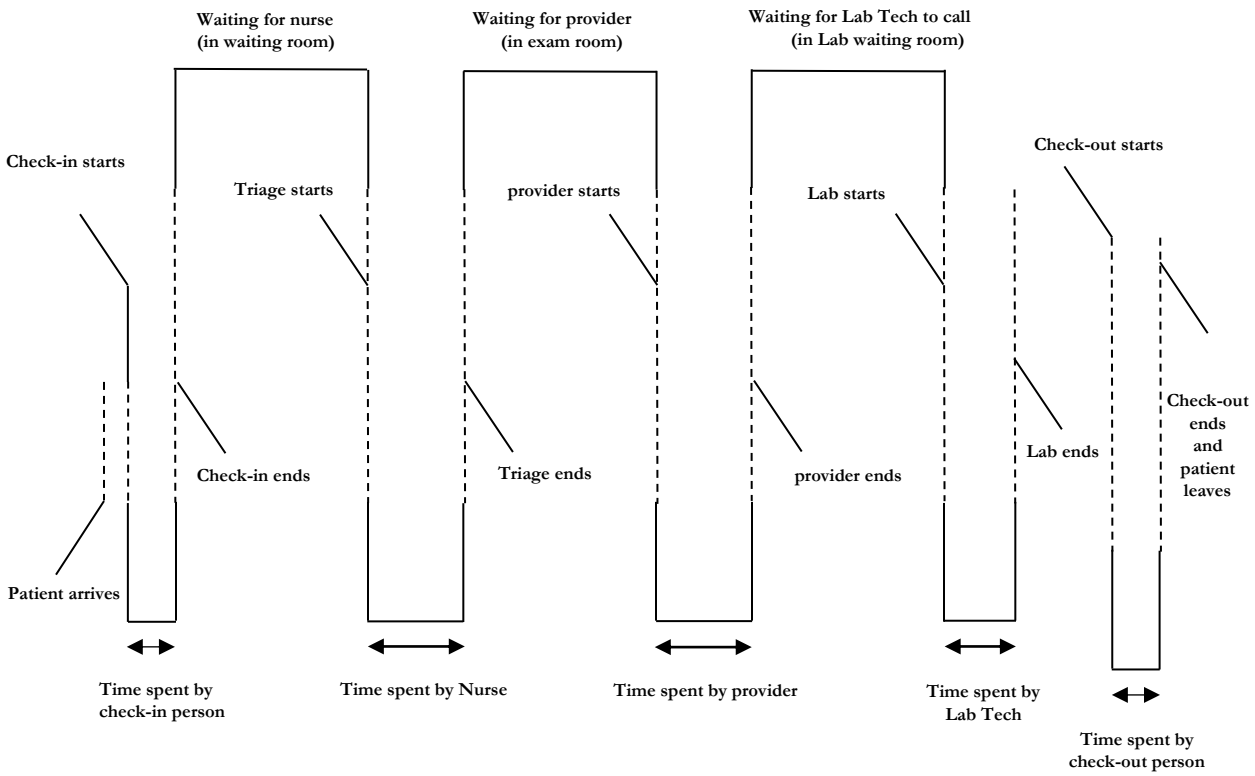

Fig 17. Flow of OBGYN patients in the clinic

Fig 18. Example of a clinic layout (https://www.ramtechmodular.com/medical-floorplans/)

3.8 Case Study

In this section, we report a case study that demonstrates how well the proposed mean-risk two -stage SMILP model approach performs in terms of the multi-category outpatient appointment scheduling for the women's clinic studied. The clinic characteristics and patient demand data used in

the case study are acquired from the literature of women's specialty clinics. The values of the parameters in the mean-risk two-stage SMILP model are selected from (Qu et al. 2013) and (Lenin et al. 2015) as well as some from preliminary numerical experiments and are denoted in Table 16. In particular, the service time durations for each patient types to visit the providers are from (Qu et al. 2013), and the service time distributions for other clinic's stations such as time spent by check-in person, nurse, lab tech, and check-out person are driven from (Lenin et al. 2015). For data and study design we refer to chapter 2 of this dissertation.

Notation	Description	Value
K	Total number of physicians available in each clinic session	2
N	Number of time slots in each clinic session	16
Δ_j	Cost of additional capacity of provider	[2000, 2000]
a_j	Number of new patients desired by provider	[10,10]
cf_i	Risk factor for patient type	[1.67, 0.4, 0.67, 1.2, 0.87, 0.67, 1]
CF_j	Risk factor for provider	[1.67, 1.67]
tlr_j	Tolerance factor of provider	[4.5, 4.5]
ρ_j	Cost of new patient type for provider	1.7
c_j	Free capacity for provider, $j \in \mathcal{R}$ over time horizon T	[2, 2]
M	A large number	4.8
$\mathcal{S}$	Set of morning/afternoon sessions over time horizon T	8
$\mathcal{g}$	Patient acceptance threshold for the first week	0.5 <= threshold< 1
$\mathfrak{I}$	Patient acceptance threshold for one month	0.2 <= threshold< 0.5
T	Time horizon	120 days
$\mathcal{f}$	Steady state	61– 100 days
λ_c	Risk/trade-off Coefficient	0.1, 0.2
α	Confidence level, (0,1]	0.1

Table. 16. Two-stage SMILP model setting parameters in the case study

3.9 Computational Results

The calculations were carried out on a Dell, 64-bit operating system, and 80 GB RAM. The solution scheme is implemented in Python 2.7.12. Gurobi is used as a solver for two-stage SMILP and SAA. In this section we present the significance of applying risk-averse two-stage SMILP approach versus risk-neutral SMILP. Definition of the base-case is the same as it is in the risk-neutral SMILP approach in chapter 2. We consider three threshold levels to check whether the daily patient flows are satisfactory. To do so, we drive the experimental results for threshold levels: λ = 50%, 65%, and 80% quantiles. The experiments designed for 50 scenarios with respect to the sample average approximation results with less gap%. The model is evaluated for different values of risk coefficients, λ_c=0.1 and 0.2, and confidence level, α=0.1.

Tables 17 and 18 present the decrease percentages of direct waiting time for different threshold levels and risk coefficients when applying risk-averse approach. The results show direct waiting time decreases up to 8% when applying risk-averse approaches compared to the risk-neutral model and up to 20% compared to the base-case. The results evaluate the waiting time for all patient types. Table 19 compares indirect waiting time decrease-% of risk-averse and risk-neutral two-stage with base-case in the first week which shows more decrease in indirect waiting time, 4.7%, at threshold level 80% for risk-averse model and 6.2% for 50% threshold level. Table 20 represents the indirect waiting time decreased up to 33% in risk-averse two-stage compared to risk-neutral approach.

Decrease in Avg. Direct Waiting Time (%)			
Risk Coefficient	**Threshold = 50%**	**Threshold = 65%**	**Threshold = 80%**
$\lambda_c = 0.1$	3.2%	2.9%	2.3%
$\lambda_c = 0.2$	8%	5.1%	2.4%

Table. 17. Advantage of risk-averse two-stage SMILP over risk-neutral two-stage SMILP for direct wait time, α =0.1

Decrease in Avg. Direct Waiting Time (%)			
Model	**Threshold=50%**	**Threshold=65%**	**Threshold = 80%**
Risk-neutral	16%	6%	3%
Risk-averse, $\lambda_c = 0.1$	16.3%	8.7%	5.3%
Risk-averse, $\lambda_c = 0.2$	20%	10.7%	5.3%

Table.18. Comparing direct wait time improvement-% of risk-averse and risk-neutral two-stage SMILP with base-case for case-2 demand scenario, α =0.1

Indirect Waiting Time				
Model	**Time-window**	**Threshold = 50%**	**Threshold = 65%**	**Threshold = 80%**
Risk-neutral	Week-1	6%	5%	1.2%
Risk-averse	Week-1	6.2%	5%	4.7%

Table.19. Comparing indirect wait time, decrease-%, of risk-averse and risk-neutral two-stage SMILP with base-case for case-2 demand scenario, Risk Coefficient, $\lambda_c = 0.2, \alpha$ =0.1

Indirect Waiting Time			
Risk Coefficient	**Threshold = 50%**	**Threshold = 65%**	**Threshold = 80%**
$\lambda_c = 0.2$	33%	31%	29%

Table.20. Advantage of risk-averse two-stage SMILP over risk-neutral two-stage SMILP for indirect wait time, α =0.1

3.10 Conclusion

In this chapter we developed mean-risk two-stage stochastic programming in which we investigate the effect of considering a risk measure in the model. We applied Conditional-Value-at-Risk (CVaR) as a risk measure for the two-stage stochastic programming model. Results from testing our models using data inspired by real-world OBGYN clinics suggest that the proposed formulations can improve patient satisfaction through reduced direct and indirect waiting times without compromising provider utilization.

In general, three directions for future research related to objective functions can be proposed. First, the linear relationship between time-based measures and their corresponding costs can be considered. Second, the Pareto approach, which provides a set of non-dominant (Pareto optimal) solutions, which is used in a few papers (Castro & Petrovic 2012) and (Qu et al. 2012). Third, risk-averse objectives can be used instead of risk-neutral objectives in order to control the variability of the target performance measures. A few optimization studies propose risk-averse objectives, such as the Markowitz mean-variance method (e.g., (Mak et al. 2015); (Qu et al. 2012)) and the Von Neumann–Morgenstern expected utility method (e.g., (Kemper et al. 2014); (Kuiper & Mandjes 2015); (LaGanga & Lawrence 2012); (Vink et al. 2015)). Other contributions are related to how often to re-execute two-stage stochastic programming, improvement to heuristic policy in call center, applying decomposition algorithm as solution approach as well as investigating meta-heuristic approaches.

CHAPTER 4: CONCLUSION AND FUTURE RESEARCH

In this dissertation, we study the application of stochastic programming in solving health care problems. In chapter two of this dissertation we mainly focus on risk-neutral two-stage stochastic programming where the objective function considers the expected value as a performance criterion. We discuss methods for improving flow through outpatient clinics considering effective appointment scheduling policies by applying two-stage Stochastic Mixed-Integer Linear Program Model (two-stage SMILP) approaches to improve patient flow metrics: direct wait time (clinic wait time), indirect wait time considering patient's no-show behavior, stochastic server, follow-up surgery appointments, and overbooking. The objective includes two models: 1) a method to optimize the (weekly) scheduling pattern for individual providers that would be updated at regular intervals (e.g., quarterly or annually) based on the type and mix of services rendered and 2) a method for dynamically scheduling patients using the weekly scheduling pattern. Scheduling will entertain the possibility of arranging multiple appointments at once (e.g., both surgery and post-surgery follow-up visits can be scheduled together for improved care).

The aim is to increase throughput per session while providing timely care, continuity of care, and overall patient satisfaction as well as equity of resource utilization. We introduced an index heuristic policy to simulate patient appointment scheduling in call center by considering patient preference date for the appointment. Finally, through clinic simulation we evaluate if the daily patient flows are satisfactory. Value of overbooking in every scheduling session was stablished through assigning different values to maximum patient criticality that provider can handle in a session. To show the advantages of two-stage programming we define base-case scenario with simulating clinic and call center using the same scenario as we design in risk-neutral two-stage SMILP approach. Our results present improvement in patient flow metrics: direct and indirect waiting time, in two-stage stochastic programming over the base-case.

In the third chapter we expand the model formulation to mean-risk two-stage stochastic programming in which we investigate the effect of considering a risk measure in the model. We focus on Conditional-Value-at-Risk (CVaR) for the risk measure as it keeps the convexity property, and one can use available solvers to solve the two-stage stochastic programming. Currently, we are working on the result preparation.

4.1 Future research

In general, three directions for future research related to objective functions can be proposed. First, the linear relationship between time-based measures and their corresponding costs can be considered. Second, the Pareto approach, which provides a set of non-dominant (Pareto optimal) solutions, which is used in a few papers (Castro & Petrovic 2012) and (Qu et al. 2012). Third, risk-averse objectives can be used instead of risk-neutral objectives in order to control the variability of the target performance measures. A few optimization studies propose risk-averse objectives, such as the Markowitz mean-variance method (e.g., (Mak et al. 2015); (Qu et al. 2012)) and the Von Neumann–Morgenstern expected utility method (e.g., (Kemper et al. 2014); (Kuiper & Mandjes 2015); (LaGanga & Lawrence 2012); (Vink et al. 2015)). Other contributions are related to how often to re-execute two-stage stochastic programming, improvement to heuristic policy in call center, applying decomposition algorithm as solution approach as well as investigating meta-heuristic approaches.

Appendix

Comparison of patient flow metrics for each patient type between Two-stage SMILP and base-case for case-2 demand scenario

Considering different threshold levels for patient flow metric distributions, Fig 19-25 compare direct wait time distributions for patient type-1 to type-7 for case-2 demand scenario between base-case and two-stage SMILP with threshold levels: λ = 50%, 65%, and 80% quantiles. The average waiting time in two-stage SMILP is less than that of base-case; improving up to 27%, 11%, 49%, 13%, 20%, 23%, and 87% for patient type-1 to type-7 in that order compare to base-case (Table 6). In Fig 19 for patient type-1, at threshold level λ = 50%, median is 20 minutes which is less than other threshold levels and the base-case. It shows 50% of population has waiting time less than 20 minutes and 25% of population in λ = 50% has less than 1 minute waiting time, and in 65% and 80% threshold levels it is less than 2 minutes while in the base-case it is less than 7 minutes.

Fig 20 shows the results for patient type-2. 25% of the population in all threshold levels has waiting time less than 5, 6, and 7 minutes while in base-case it is less than 9 minutes. In addition, 75% of population with less than 14 and 15 minutes waiting time in two-stage SMILP shows robust results. For patient typ-3, Fig 21 depicts the median in all threshold levels is less than that in the base-case; 9 minutes versus 14.5 minutes. 25% of the population in all threshold levels has no waiting time whereas in the base-case, 25% of the population has waiting time less than 8 minutes. Moreover, 75% of population has waiting time less than 17, 18, and 19 minutes for all threshold levels in two-stage SMILP compared to the 24 minutes in base-case. Fig 22 for patient type-4 shows 25% of the population in threshold level: λ = 50% has waiting time less than 10.5 minutes while 25% of the population has waiting time less than 13 minutes in the base-case. In Fig 23, median in all threshold levels for patient type-5 is less than that in the base-case and 25% of the population in all threshold levels has less than 7 and 9.5 minutes of waiting time comparing to that in the base-case with less than

11 minutes. In addition, 75% of population has waiting time at most 19 minutes in two-stage SMILP comparing to 21 minutes in the base-case.

Fig 24 denotes up to 8 minutes waiting time in two-stage SMILP and 12 minutes in base-case covering 25% of the population of patient type-6. Moreover, 75% of population has up to 19 minutes waiting time for all threshold levels in two-stage SMILP compared to 21 minutes in the base-case. The graph of waiting time for patient type-7 in Fig 25 shows less median in all threshold levels compared to the base-case. Also, 25% of the population has up to 3 minutes waiting time in two-stage SMILP verses 12 minutes in the base-case and 75% of population has at most 18 minutes of waiting time in two-stage SMILP compares to 20 minutes for the base-case.

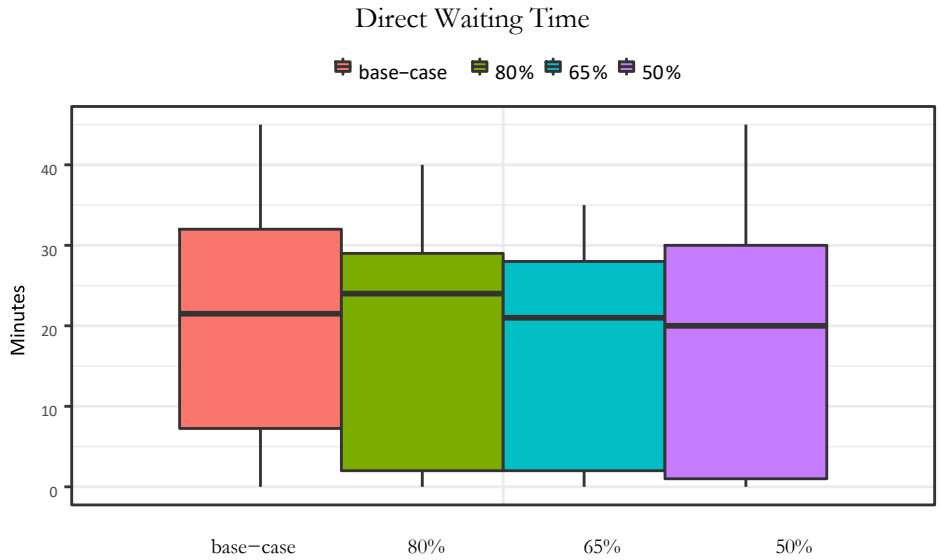

Fig 19. Direct wait time distribution for patient type-1 for case-2 demand scenario, comparing base-case and Two-stage SMILP with threshold level: λ = 50%, 65%, and 80%

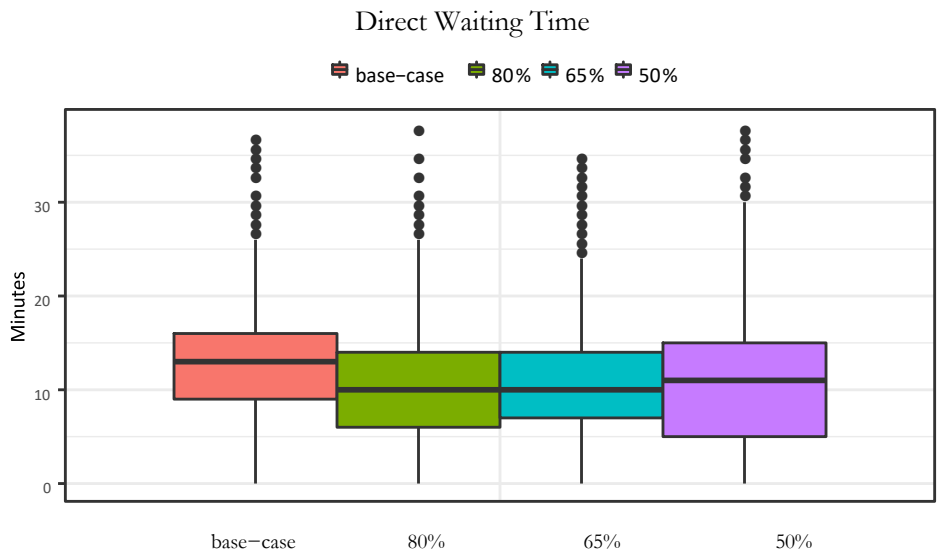

Fig 20. Direct wait time distribution for patient type-2 for case-2 demand scenario, comparing base-case and Two-stage SMILP with threshold level: λ = 50%, 65%, and 80%

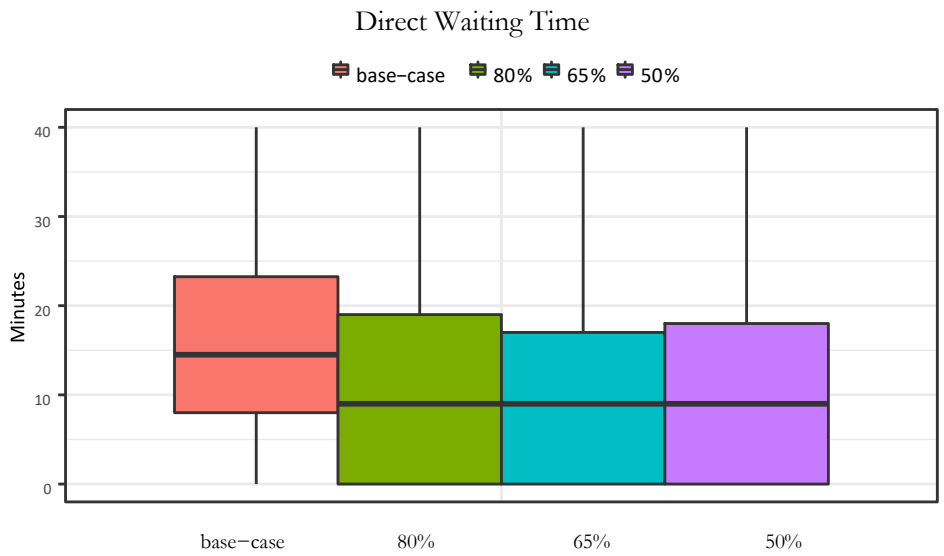

Fig 21. Direct wait time distribution for patient type-3 for case-2 demand scenario, comparing base-case and Two-stage SMILP with threshold level: λ = 50%, 65%, and 80%

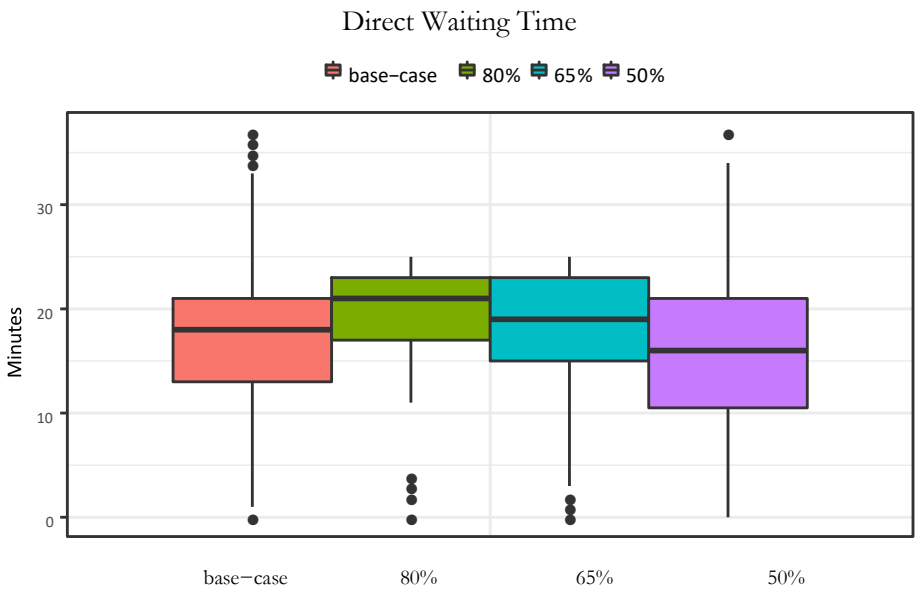

Fig 22. Direct wait time distribution for patient type-4 for case-2 demand scenario, comparing base-case and Two-stage SMILP with threshold level: λ = 50%, 65%, and 80%

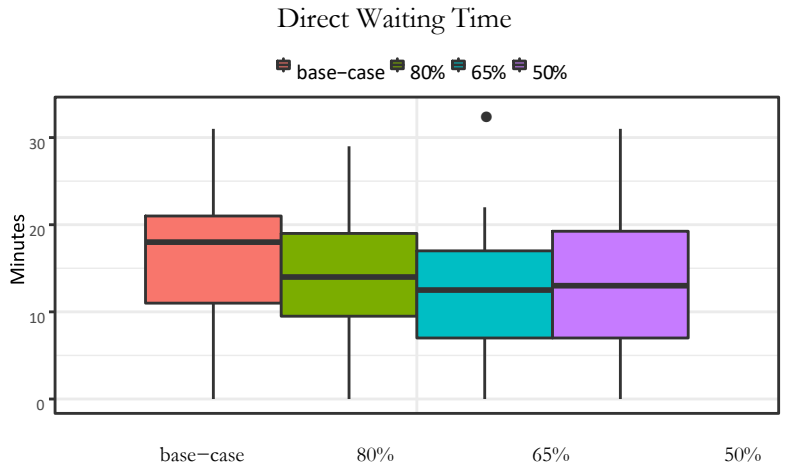

Fig 23. Direct wait time distribution for patient type-5 for case-2 demand scenario, comparing base-case and Two-stage SMILP with threshold level: λ = 50%, 65%, and 80%

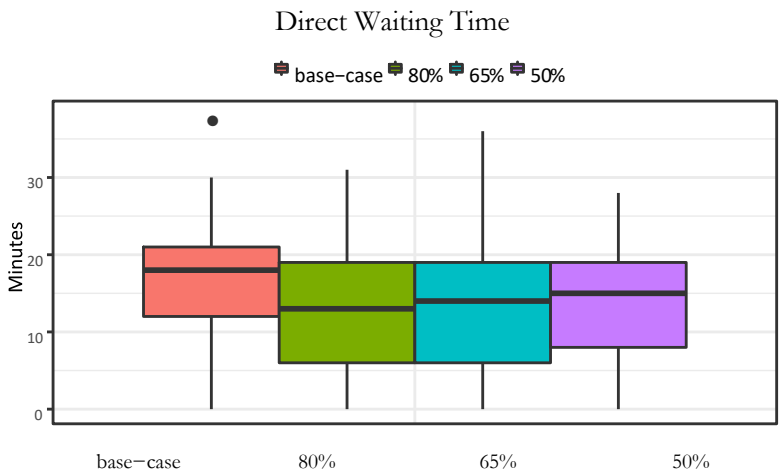

Fig 24. Direct wait time distribution for patient type-6 for case-2 demand scenario, comparing base-case and Two-stage SMILP with threshold level: λ = 50%, 65%, and 80%

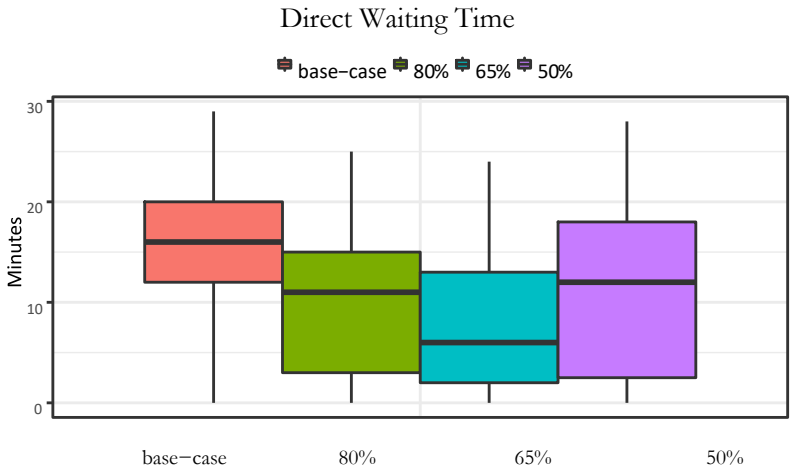

Fig 25. Direct wait time distribution for patient type-7 for case-2 demand scenario, comparing base-case and Two-stage SMILP with threshold level: λ = 50%, 65%, and 80%

Trade-off between direct wait time and indirect wait time for each patient type between Two-stage SMILP and base-case for case-2 demand scenario

Fig (26-32) represent the trade-off between direct wait time and indirect wait time distributions for two-stage stochastic programming (SMILP) with different threshold levels λ = 50%, 65%, and 80% quantiles and the base-case. The average waiting time is higher in the base-case, comparing to other threshold levels. Therefore, we expect the crowded clinic days for indirect wait time in the beginning of the time horizon comparing to other threshold levels. Base-case shows weaker result

compares to two-stage stochastic programming as it has busy clinic days at the first two weeks and some appointments for weeks 3 and 4 which shows high waiting time in the system.

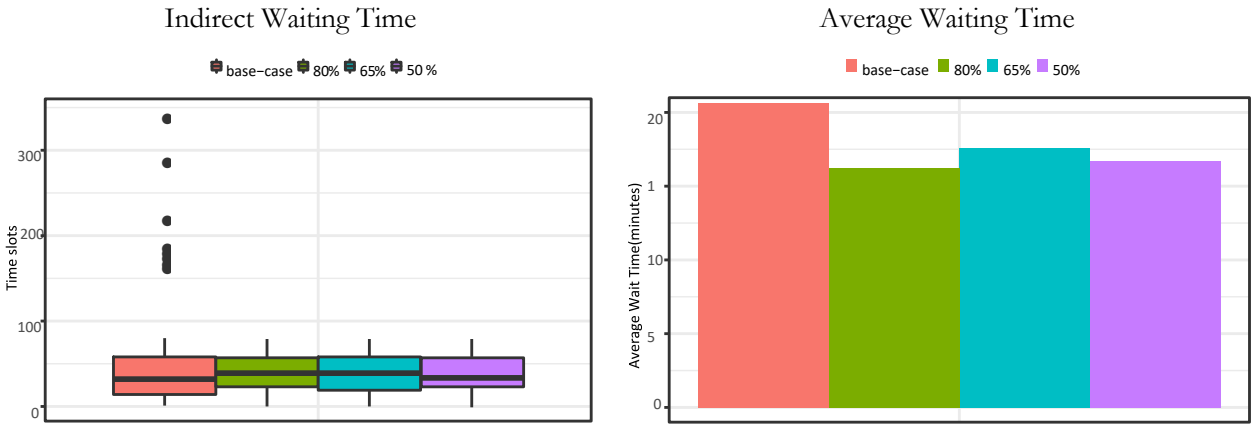

Fig 26. Trade-off between direct wait time and indirect wait time distributions between Two-stage SMILP with threshold levels: λ = 50%, 65%, and 80% quantile and base-case for patient type-1 for case-2 demand scenario

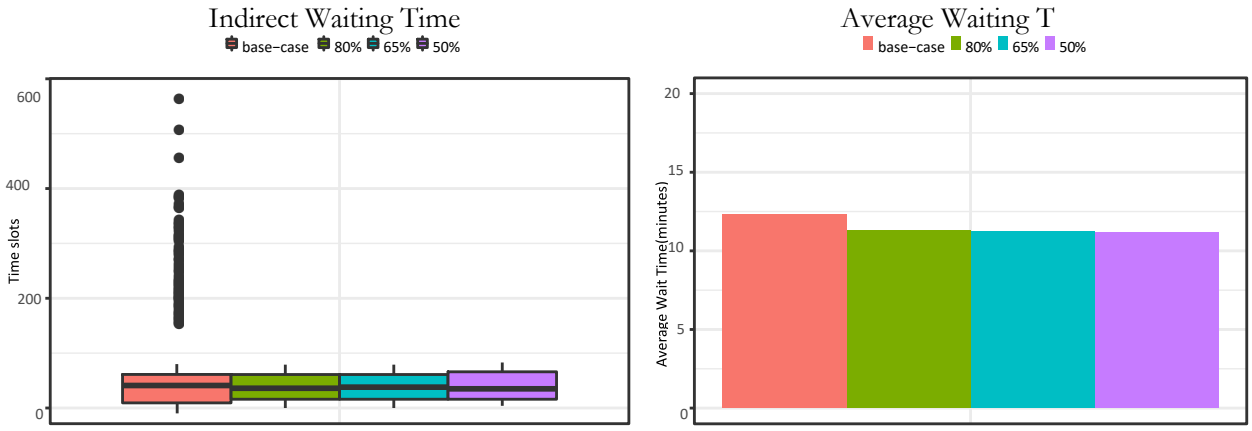

Fig 27. Trade-off between direct wait time and indirect wait time distributions between Two-stage SMILP with threshold levels: λ = 50%, 65%, and 80% quantile and base-case for patient type-2 for case-2 demand scenario

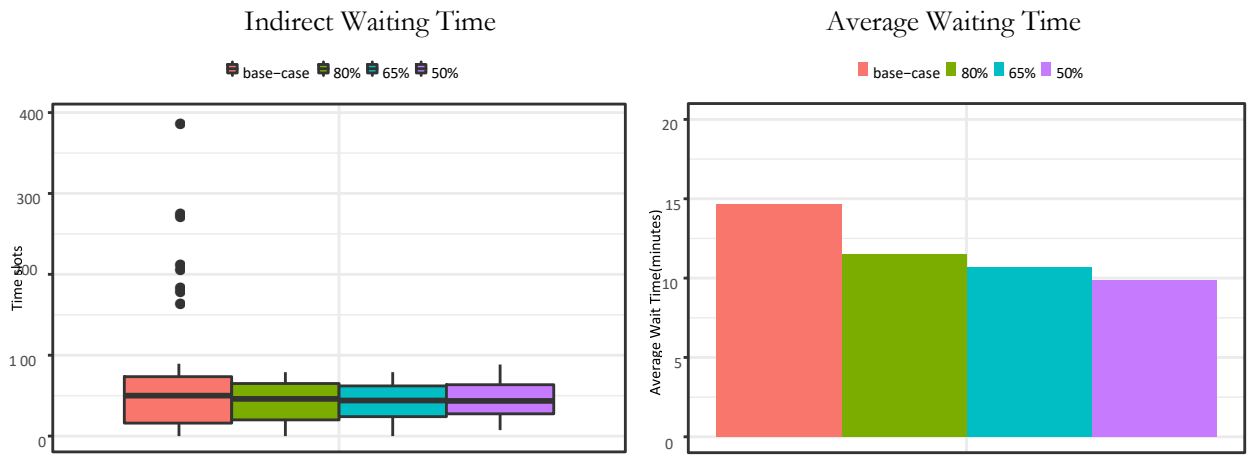

Fig 28. Trade-off between direct wait time and indirect wait time distributions between Two-stage SMILP with threshold levels: λ = 50%, 65%, and 80% quantile and base-case for patient type-3 for case-2 demand scenario

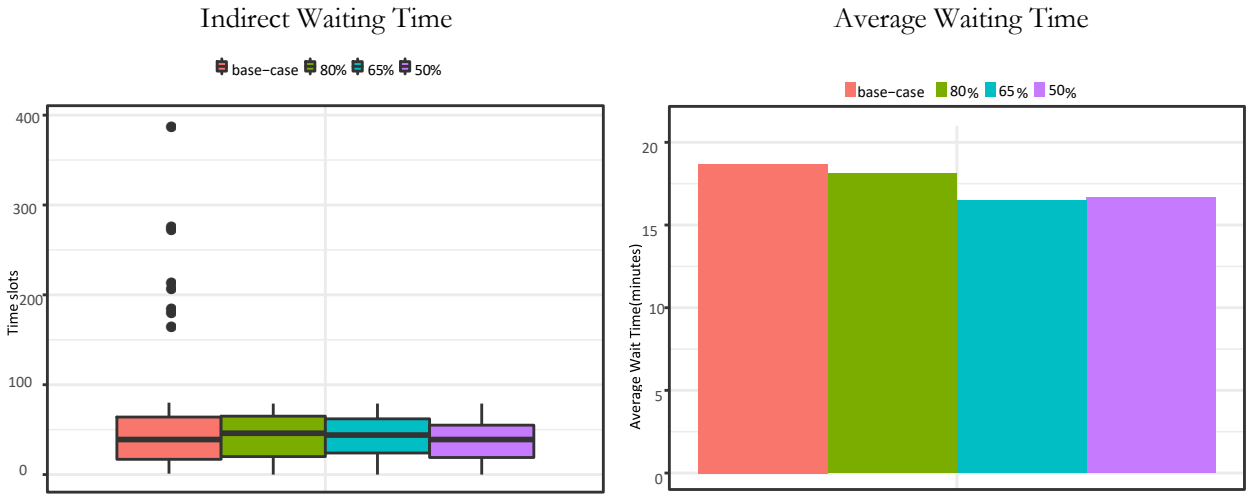

Fig 29. Trade-off between direct wait time and indirect wait time distributions between Two-stage SMILP with threshold levels: λ = 50%, 65%, and 80% quantile and base-case for patient type-4 for case-2 demand scenario

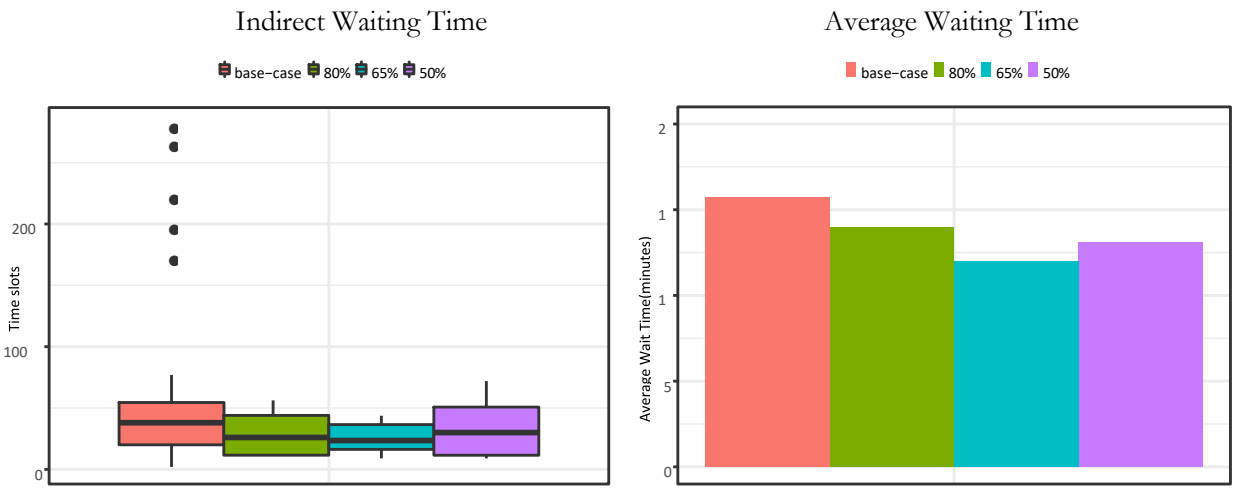

Fig 30. Trade-off between direct wait time and indirect wait time distributions between Two-stage SMILP with threshold levels: λ = 50%, 65%, and 80% quantile and base-case for patient type-5 for case-2 demand scenario

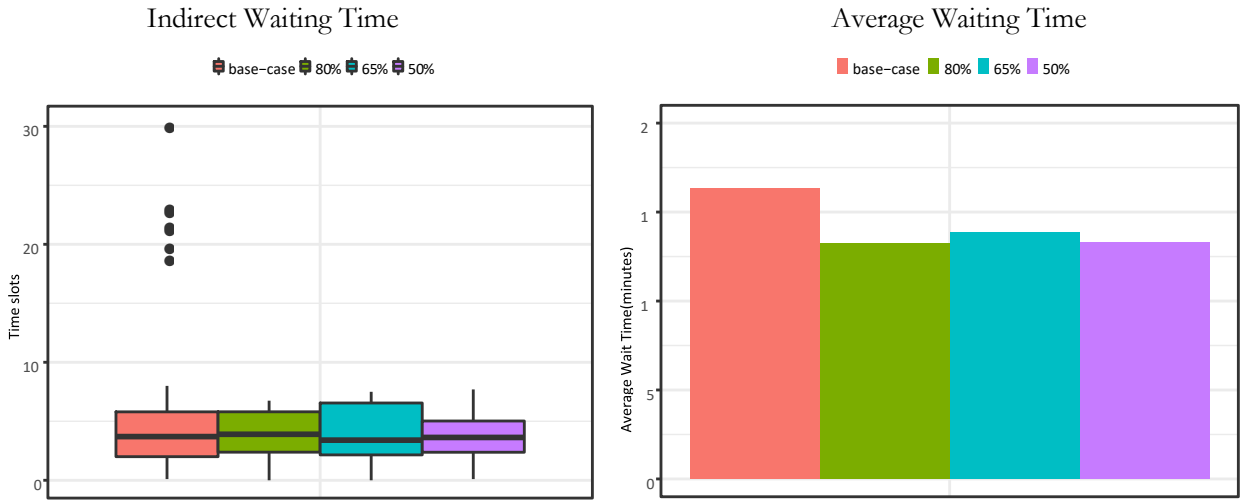

Fig 31. Trade-off between direct wait time and indirect wait time distributions between Two-stage SMILP with threshold levels: λ = 50%, 65%, and 80% quantile and base-case for patient type-6 for case-2 demand scenario

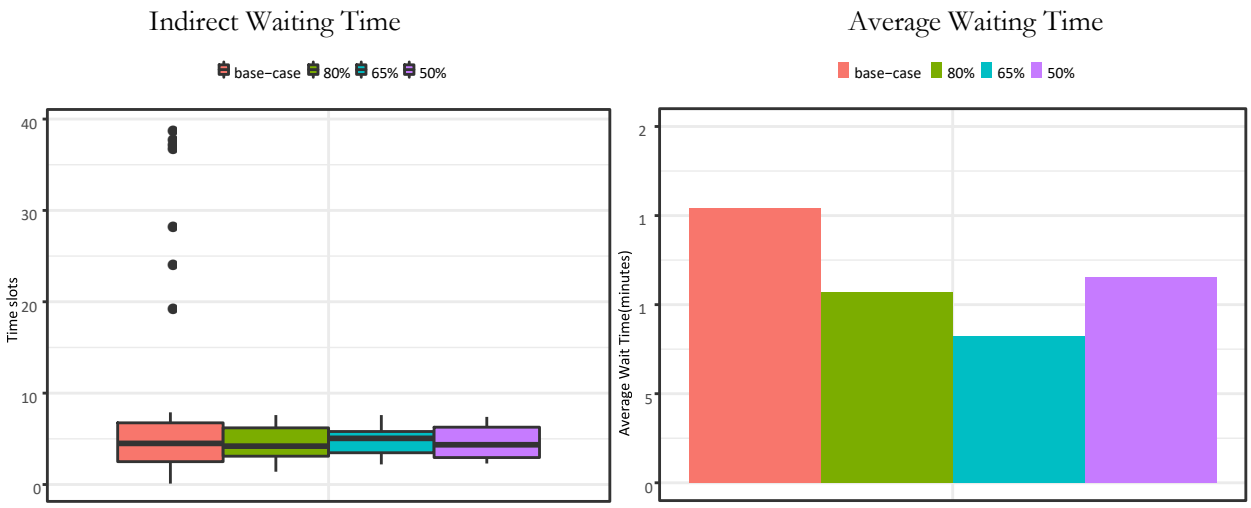

Fig 32. Trade-off between direct wait time and indirect wait time distributions between Two-stage SMILP with threshold levels: λ = 50%, 65%, and 80% quantile and base-case for patient type-7 for case-2 demand scenario

References

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

Ahmed, S., 2006. Convexity and decomposition of mean-risk stochastic programs. Mathematical Programming, 106(3), pp.433–446.

Ahmed, S., 2004. Mean-risk objectives in stochastic programming. NSF Design and Manufacturing Grantees Conference, (3), pp.2003–2005.

Anon, 2001. Crossing the quality chasm: a new health system for the 21st century.

Asch, S.M. et al., 2006. Who Is at Greatest Risk for Receiving Poor-Quality Health Care? New England Journal of Medicine, 354(11), pp.1147–1156.

Balasubramanian, J. & Grossmann, I., 2003. Scheduling optimization under uncertainty—an alternative approach. Computers and Chemical Engineering, 27, pp.469–490.

Berg, B.P. et al., 2014. Optimal booking and scheduling in outpatient procedure centers. Computers and Operations Research, 50, pp.24–37.

Birge, J. & Louveaux, F., 1997. Introduction to Stochastic Programming. Springer Series in Operations Research, p.421.

Camacho, F. et al., The relationship between patient's perceived waiting time and office-based practice satisfaction. North Carolina medical journal, 67(6), pp.409–13.

Cartwright, A. et al., 1992. Outpatients and their doctors : a study of patients, potential patients, general practitioners and hospital doctors, HMSO.

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

Cayirli, T. & Veral, E., 2003. Outpatient Scheduling in Health Care: a Review of Literature. Production and Operations Management, 12(4), pp.519–549.

Daniels, R.L. et al., 1995. Robust Scheduling to Hedge Against Processing Time Uncertainty in Single-stage Production. , 41(2), pp.363–376.

De, P., Ghosh, J.B. & Wells, C.E., 1992. Expectation-variance analyss of job sequences under processing time uncertainty. International Journal of Production Economics, 28(3), pp.289–297.

DeFife, J. a et al., 2010. Psychotherapy appointment no-shows: rates and reasons. Psychotherapy Theory, Research, Practice, Training, 47(3), pp.413–417.

Dreiher, J. et al., 2008. Nonattendance in obstetrics and gynecology patients. Gynecologic and Obstetric Investigation, 66(1), pp.40–43.

Duffie, D. & Pan, J., 1997. An Overview of Value at Risk. The Journal of Derivatives, 4(3), pp.7–49.

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

Erdogan, S.A., Gose, A. & Denton, B.T., 2015. On-line Appointment Sequencing and Scheduling. IIE Transactions, 8830(May 2015), pp.00–00.

Fábián, C.I., 2008. Handling CVaR objectives and constraints in two-stage stochastic models. European Journal of Operational Research, 191(3), pp.888–911.

Feldman, J. et al., 2014. Appointment Scheduling Under Patient Preference and No-Show Behavior. Operations Research, 62(4), pp.794–811.

Ferguson, A.R. & Dantzig, G.B., 1956. The Allocation of Aircraft to Routes—An Example of Linear Programming Under Uncertain Demand. Management Science, 3(1), pp.45–73.

Festinger, D.S. et al., 2002. From telephone to office: intake attendance as a function of appointment delay. Addictive behaviors, 27(1), pp.131–7.

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

Hawkins, M. & Irving, S.D., 2017. 2017 Survey of Physician Appointment Wait Times. , 75063(800).

Hill, C.J. & Joonas, K., 2005. The Impact of Unacceptable Wait Time on Health Care Patients' Attitudes and Actions. Health Marketing Quarterly, 23(2), pp.69–87.

Huang, X.-M., 1994. Patient Attitude towards Waiting in an Outpatient Clinic and its Applications. Health Services Management Research, 7(1), pp.2–8.

Kemper, B., Klaassen, C.A.J. & Mandjes, M., 2014. Optimized appointment scheduling. European Journal of Operational Research, 239(1), pp.243–255.

Kohn, L.T., Corrigan, J.M. & Donaldson, M.S., 2000. To err is human: building a safer health system

Kouvelis, P., Daniels, R.L. & Vairaktarakis, G., 2000. Robust scheduling of a two-machine flow shop with uncertain processing times. IIE Transactions (Institute of Industrial Engineers), 32(5), pp.421–432.

Krokhmal, P., Zabarankin, M. & Uryasev, S., 2011. Modeling and optimization of risk. Surveys in Operations Research and Management Science, 16(2), pp.49–66.

Kuiper, A. & Mandjes, M., 2015. Appointment scheduling in tandem-type service systems. Omega (United Kingdom), 57, pp.145–156.

Laganga, L.R. & Lawrence, S.R., 2007. Clinic overbooking to improve patient access and increase provider productivity. Decision Sciences, 38(2), pp.251–276.

LaGanga, L.R. & Lawrence, S.R., 2012. Appointment Overbooking in Health Care Clinics to Improve Patient Service and Clinic Performance. Production and Operations Management, 21(5), pp.874–888.

Lenin, R.B. et al., 2015. Optimizing appointment template and number of staff of an OB/GYN clinic--micro and macro simulation analyses. BMC health services research, 15(1), p.387.

Liu, N., Ziya, S. & Kulkarni, V.G., 2010. Dynamic Scheduling of Outpatient Appointments Under Patient No-Shows and Cancellations. Manufacturing & Service Operations Management, 12(2), pp.347–364.

Macharia, W.M. et al., 1992. An overview of interventions to improve compliance with appointment keeping for medical services. JAMA : the journal of the American Medical Association, 267(13), pp.1813–7.

Mak, H.-Y., Rong, Y. & Zhang, J., 2015. Appointment Scheduling with Limited Distributional Information. Management Science, 61(2), pp.316–334.

Mak, W.K., Morton, D.P. & Wood, R.K., 1999. Monte Carlo bounding techniques for determining solution quality in stochastic programs. Operations Research Letters, 24(1), pp.47–56.

Markowitz, H., 1952. Portfolio Selection. The Journal of Finance, 7(1), pp.77–91.

McCarthy, K., McGee, H.M. & O'Boyle, C.A., 2000. Outpatient clinic waiting times and non-attendance as indicators of quality. Psychology, Health & Medicine, 5(3), pp.287–293.

Morgan, J.P., 1994. RiskMetrics — Technical Document.

Muthuraman, K. & Lawley, M., 2008. A stochastic overbooking model for outpatient clinical scheduling with no-shows. IIE Transactions, 40(9), pp.820–837.

Norkin, V., Pflug, G. & Ruszczyński, A., 1998. A branch and bound method for stochastic global optimization. Mathematical programming, 83(98), pp.425–450.

Noyan, N., 2012. Risk-averse two-stage stochastic programming with an application to disaster management. Computers & Operations Research, 39(3), pp.541–559.

Ogryczak, W. & Ruszczyski, A., 2002. Dual stochastic dominance and related mean-risk models. SIAM Journal on Optimization, 13(1), pp.60–78.

Patrick, J., Puterman, M.L. & Queyranne, M., 2008. Dynamic Multipriority Patient Scheduling for a Diagnostic Resource. Operations Research, 56(6), pp.1507–1525.

Poses, R.M., 2003. A cautionary tale: the dysfunction of American health care. European journal of internal medicine, 14(2), pp.123–130.

Qi, J., 2017. Mitigating Delays and Unfairness in Appointment Systems. Management Science, 63(2),

pp.566–583.

Qu, X. et al., 2013. A two-phase approach to scheduling multi-category outpatient appointments - A case study of a women's clinic. Health Care Management Science, 16(3), pp.197–216.

Qu, X., Rardin, R.L. & Williams, J.A.S., 2012. A mean–variance model to optimize the fixed versus open appointment percentages in open access scheduling systems. Decision Support Systems, 53(3), pp.554–564.

Rau, J., 2011. Medicare To Begin Basing Hospital Payments On Patient-Satisfaction Scores. Kaiser Health News.

RE, H. (2006), Why Innovation in Health Care Is So Hard.

Sabuncuoglu, I. & Bayiz, M., 2000. Analysis of reactive scheduling problems in a job shop environment. European Journal of Operational Research, 126(3), pp.567–586.

Santoso, T. et al., 2005. A stochastic programming approach for supply chain network design under uncertainty. European Journal of Operational Research, 167(1), pp.96–115.

Sarin, S.C., Sherali, H.D. & Liao, L., 2014. Minimizing conditional-value-at-risk for stochastic scheduling problems. Journal of Scheduling, 17(1), pp.5–15.

Schultz, R. & Tiedemann, S., 2006. Conditional value-at-risk in stochastic programs with mixed-integer recourse. Mathematical Programming, 105(2–3), pp.365–386.

Schultz, R. & Tiedemann, S., 2003. Risk aversion via excess probabilities in stochastic programs with mixed-integer recourse. SIAM Journal on Optimization, 14(1), pp.115–138.

Schuster, M.A., McGlynn, E.A. & Brook, R.H., 1998. How good is the quality of health care in the United States? The Milbank quarterly, 76(4), pp.517–63, 509.

Schutz, P., Tomasgard, A. & Ahmed, S., 2009. Supply chain design under uncertainty using sample average approximation and dual decomposition. European Journal of Operational Research, 199(2), pp.409–419.

Skutella, M. & Uetz, M., 2005. Stochastic Machine Scheduling with Precedence Constraints. SIAM Journal on Computing, 34(4), pp.788–802.

Toh, L.S. & Sern, C.W., 2011. Patient waiting time as a key performance indicator at orthodontic specialist clinics in selangor. Malaysian Journal of Public Health Medicine, 11(1), pp.60–69.

Verweij, B. et al., 2003. The Sample Average Approximation Method Applied to Stochastic Routing Problems: A Computational Study. Computational Optimization and Applications, 24(2–3), pp.289–333.

Vink, W. et al., 2015. Optimal appointment scheduling in continuous time: The lag order approximation method. European Journal of Operational Research, 240(1), pp.213–219.

Wenger, N.S. et al., 2003. The quality of medical care provided to vulnerable community-dwelling older patients. Annals of internal medicine, 139(9), pp.740–7.

Zacharias, C. & Armony, M., 2016. Joint Panel Sizing and Appointment Scheduling in Outpatient Care. Management Science, (December), pp.1–35.

Zacharias, C. & Pinedo, M., 2014. Appointment Scheduling with No-Shows and Overbooking. Production & Operations Management, 23(5), pp.788–801.

Zeng, B. et al., 2010. Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities. Annals of Operations Research, 178(1), pp.121–144.

Zenios, S. a., Chertow, G.M. & Wein, L.M., 2000. Dynamic Allocation of Kidneys to Candidates on the Transplant Waiting List. Operations Research, 48(4), pp.549–569.

ABSTRACT

MANAGING OPERATIONAL EFFICIENCY AND HEALTH OUTCOMES AT OUTPATIENT CLINICS THROUGH EFFECTIVE SCHEDULING

by

SAMIRA FAZEL ANVARYAZDI

August 2018

Primary Major Advisor: Dr. Saravanan Venkatachalam

Co-Major Advisor: Dr. Ratna Babu Chinnam

Major: Industrial Engineering

Degree: Doctor of Philosophy

A variety of studies have documented the substantial deficiencies in the quality of health care delivered across the United States. Attempts to reform the United States health care system in the 1980s and 1990s were inspired 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, access, quality, and costs are still challenges, and medical professionals are increasingly dissatisfied.

In recent years, appointment scheduling in outpatient clinics has attracted much attention in health care delivery systems. Increase in demand for health care services as well as health care costs are the most important reasons and motivations for health care decision makers to improve health care systems. The goals of health care systems include patient satisfaction as well as system utilization. Historically, less attention was given to patient satisfaction compared to system utilization and conveniences of care providers. Recently, health care systems have started setting goals regarding patient satisfaction and improving the performance of the health system by providing timely and appropriate health care delivery.

In this study we discuss methods for improving patient flow through outpatient clinics considering effective appointment scheduling policies by applying two-stage Stochastic Mixed-Integer Linear Program Model (two-stage SMILP) approaches. Goal is to improve the following patient flow metrics: direct wait time (clinic wait time) and indirect wait time considering patient's no-show behavior, stochastic server, follow-up surgery appointments, and overbooking. The research seeks to develop two models: 1) a method to optimize the (weekly) scheduling pattern for individual providers that would be updated at regular intervals (e.g., quarterly or annually) based on the type and mix of services rendered and 2) a method for dynamically scheduling patients using the weekly scheduling pattern. Scheduling templates will entertain the possibility of arranging multiple appointments at once. The aim is to increase throughput per session while providing timely care, continuity of care, and overall patient satisfaction as well as equity of resource utilization. First, we use risk-neutral two-stage stochastic programming model where the objective function considers the expected value as a performance criterion in the selection of random variables like total waiting times and next, we expand the model formulation to mean-risk two-stage stochastic programming in which we investigate the effect of considering a risk measure in the model. We apply Conditional-Value-at-Risk (CVaR) as a risk measure for the two-stage stochastic programming model. Results from testing our models using data inspired by real-world OBGYN clinics suggest that the proposed formulations can improve patient satisfaction through reduced direct and indirect waiting times without compromising provider utilization.

AUTOBIOGRAPHICAL STATEMENT

Samira Fazel Anvaryazdi received her BSc degree in applied mathematics from Isfahan Payam-Noor University. She received her first MSc degree in pure mathematics from Isfahan University and her second MSc degree in mathematical statistics from Wayne State University. She received her PhD degree in industrial and systems engineering at Wayne State University. Her research interests include scheduling, stochastic programming, robust optimization, and compartmental models in healthcare systems engineering. During her study at Wayne State University she presented her research at INFORMS and Graduate and Postdoctoral Research Symposium at Wayne State University. Moreover, she was a co-chair of the 2016 & 2017 Graduate Research Symposium in Industrial & Systems Engineering Department.

Throughout her academic career, she has accumulated nearly ten years of teaching experience as an instructor teaching mathematics and statistics. She also regularly follows teaching and learning events at Office of Teaching & Learning (OTL) and NIH B.E.S.T. workshops as well as OTL Pedagogy Journal Club at Wayne State University and Lawrence Technological University. She won the 2017 GEOC teaching award at Wayne State University. In 2018, she won Integrating Curriculum with Entrepreneurial Mindset (ICE) - KEEN ICE Award and joined a KEEN Innovative Teaching (KIT) faculty member and become part of a unique cohort of faculty who are committed to improving engineering education.

Her papers are under review in venues such as IISE Transactions on Healthcare Systems Engineering and Healthcare management. She is a student member of AMS and INFORMS. Following graduation, Samira will join the Louisiana Tech University as a Visiting Assistant Professor of Statistics and Industrial Engineering.