Stochastic Optimization Approaches for Outpatient Appointment Scheduling under Uncertainty

by

Karmel S. Shehadeh

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in the University of Michigan
2019

Doctoral Committee:

Professor Amy E.M. Cohn, Co-Chair
Assistant Professor Ruiwei Jiang, Co-Chair
Professor Marina A. Epelman
Professor Anne E. Sales

Karmel S. Shehadeh

ksheha@umich.edu

ORCID iD: 0000-0001-7842-0951

Dedication

I dedicate this dissertation to all the girls and boys in the world. If you are reading this, I want you to be proud of who you are. If you have a dream; believe it, fight for it, and work hard to achieve it. They say that there's a discipline for reaching your dreams. And it's not about the number of times you fall down or get rejected by others, but rather about how you stand up, bravely, for what you believe in, how you follow your heart, and how you keep going in this life with a cheerful spirit. Find a mentor. If you look around you, you will find many. Trust them and work with them in finding what is best for you. You can change this world into a better place; you just have to believe in the light within your heart and spread love wherever you go. Take this advice to guide you through life...to create your own story, magic, and dreams.

I also dedicate this dissertation and all the associated publications to the memory of Prof. Shabbir Ahmed (who contributed significantly to our field), all the people who refuse to give up on others, and, of course, to my incredibly amazing, loving, and supportive mother, Jumana, and big sister, Zoya.

Acknowledgments

My Ph.D. journey at the University of Michigan is, by far, the most rewarding, enjoyable, and unforgettable experience I've ever had. I have not traveled in a vacuum in this journey. My Ph.D. life and this dissertation have been kept on track with the love, support, and encouragement of many wonderful individuals and prestigious institutions to whom I owe my success and the completion of this dissertation. This acknowledgment demonstrates part of my deep appreciations to everyone and a promise to leverage what I learned from each in helping the future generations reach their dreams.

Foremost, I would like to express my sincere gratitude to my two advisors, Professor Amy Cohn and Professor Ruiwei Jiang. Thank you so much for treating me as a colleague, allowing me to be an independent researcher, challenging me to look at research problems from a variety of perspectives, trusting and growing my technical and practical research ideas and skills, and providing me with the confidence to voice my ideas and concerns. I am indeed fortunate because it is rare to find supportive advisors and colleagues like you who always find fun in research and be there to listen to and help in little problems and roadblocks that unavoidably crop up in the course of conducting research. You've taught me more than I could ever give you credit for here and shown me how to be a successful and excellent teacher, scholar, mentor, friend, and person. I feel so honored not only because I have worked with you but also that I have met and known you. I leave you with great sadness but also with anticipation for future collaboration and research adventures.

Besides my advisors, I want to express my special appreciation and thanks to my committee member Professor Marina A. Epelman, who have taught me, both consciously and unconsciously, how an influential and successful mentor, scholar, teacher, and person should be. I appreciate all her contributions of time, idea, and support to make my research and my Ph.D. life productive, rewarding, and stimulating. The joy and enthusiasm she has for research, teaching, and Apple products was contagious and motivational for me, even during tough times in the Ph.D. pursuit. I very much enjoyed working closely with her on a technical paper (Chapter 2), and I am thankful for her generous support during my time here. I leave her with great sadness but also with anticipation for future collaboration and research adventures.

I am also thankfull for my committee member Professor Anne Sales, our clinical collaborators Dr. Sameer Saini and Dr. Jacob Kurlander from Michigan Medicine, and everyone from the Gastroenterology Learning Community. Your support and insights contributed to this dissertation in so many ways. I am looking forward to continuing our collaborations to achieve better care in Michigan Medicine and other health systems.

I am deeply thankful for financial the support of the Seth Bonder foundation, Department of Industrial and Operations Engineering (IOE), and Center for Healthcare Engineering and Patient Safety (CHEPS). I am particularly thankful for Mrs. Merrill Bonder for being a great source of academic and personal support, an essential part of my story, and for always giving me a happy memory to remember.

The IOE and CHEPS families are amazing with supportive faculty, staff, and students that my appreciation for them needs pages to write. Rebekah Smith and Catherine Boblitt thank you for being more than amazing sisters, for being always there for me, and for all the life and academic advice. Matt Irelan, you are a great graduate program advisor, and I will always be thankful for all that you did to help me during my Ph.D. journey. Tina, thank you for being an extraordinary Michigan mom for me, for taking care of me, and for making sure that I am eating well (I loved all your food). Rod Capps and Christopher Konrad thanks for being amazing big brothers, helping me in computer, technical, and life issues. Rod, I promise you to be a good guitar player and to grow my love for music further. IOE and CHEPS staff, Akosua, Billy, Gene, Gwen, Julia, Liz, Mint, Tom, Valerie, and Wanda, thank you so much for all that you do to help us in paperwork, teaching, preparing for conferences, and more.

Prof. Mark Daskin, thank you for going beyond and above to support me during my time here and my service as a graduate student coordinator. In particular, Ill be forever grateful for you for choosing me to teach IOE 543-Scheduling Theory and Algorithms as a sole instructor, which was an extraordinary academic opportunity and experience. Prof. Jon Lee, thank you so much for supporting me both academically and personally. I enjoyed every meeting with you, and I am looking forward to collaborating with you on a legendary integer programming research problems. Prof. Brian Denton, thank you for growing my interest in and love for scheduling and for all your support while teaching IOE 543 and serving as the graduate student coordinator. The IOE family are lucky to have you as our department chair. IOE faculty, thank you all for supporting my journey and for being a great source of inspiration and encouragement.

A special thanks go to my absolutely fantastic Cohort and other IOE friends. I learned a lot from each one of you, and I very much enjoyed this journey with you. I will always remember our late night research and course work together, our monthly dinner and gathering, and our adventures in Ann Arbor. Alejandro, Ece, Elnaz, Emily, Donald, Gian, Ismael, Maya, Niusha, Kayse, Victor, Wesley, Wilmer, Xiangkun, Yiling, Yidu, Young Chea, and Yuanyuan, I feel full of honor to have such brilliant scholars and supportive friends like you.

Next thank you to all of my CHEPS colleagues and SCOPES teammates. I very much enjoyed working with each one of you. William Yang, thank you for being a fantastic mentee. I cannot wait to hear all about your accomplishments during your upcoming Ph.D. journey and afterward. My IOE 543 students, thank you for making IOE 543 an intellectually rewarding experience. I left every lecture on a high for the rest of the day.

I also want to thank everyone from the System Science and Industrial Engineering family at Binghamton University (BU). In particular, Prof. Chun-An Chou and Prof. Muhammed Khasawneh thank you for supporting me during my time at BU and during my Ph.D. study at Michigan. I am also thankful for all my school teachers who grew my love for math and science at a young age.

Of course, no acknowledgments would be complete without giving a huge thanks to my parents, Jumana and Sami. First of all, thanks for the ‘smart genes’ you passed on to me and for giving me the life that every child deserves. I would not be who I am today without your endless love and support, and there are not enough words to describe how thankful I am to you. You instilled admirable qualities in me, gave me an optimal foundation with which to meet life, and taught me about love, friendship, hard work, self-respect, acceptance, persistence, independence, and more. Mom, especially, you are a great role model of success, resilience, strength, and character. Thank you both for supporting my love for math at a young age and for sending me to the best schools to achieve my academic dream. The way you support my dreams and the pride and love that I see in your eyes when you look at me mean the absolute world to me and makes me want to be just like you with my children.

My siblings Zoya, Rand, Shams, and Ali, you are my strength in this world, and you are the most precious people in my life. I want you all to know that I am so thankful for all the things you say and do. The “I love you” text that each one of you sends me every morning, truly make my day. There is nothing better than knowing that I have your unconditional and consistent love and support. Zoya, thank you for being an extraordinary big sister and for selflessly encouraging me to explore new directions in life and seek my own destiny. I couldn’t have done it without your endless love and support.

Last but not least, I want to thank all my extended family and friends. To my uncle and aunt, Omar and Darla Abu-Shanab, thank you for taking me under your wings from the very first day of my graduate studies in the United States. You are both amazing, and Ill be forever grateful for the immeasurable support you gave me during this journey. Thank you for giving me a home to come back to anytime I needed one. You always expressed how proud you are of me, came to my special events, and pushed me to achieve my dreams. I couldnt have done it without your love and the love of your children: Kylee, Lutfi, Aziz, and Humzah. Of course, I must never forget Grandma Shifa, Uncles Ahmed and Abdul and all of my aunties and cousins. Thank you. Hala and Kays Sawaged, my dearest friends, thank you for being always there for me and for your continuous support during this journey. My friends, Arwa, Bassim, Dima, Maysa, and Rahmeh, each one of you kept me so strong and have contributed to my academic success in so many ways.

I could not have done any of this without your love and generous support. Thank you ALL from the bottom of my heart.

Karmel S. Shehadeh

TABLE OF CONTENTS

Chapter

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF ABBREVIATIONS

AO appointment order

ARG approximate (statistical) relative gap

CRC colorectal cancer

DR distributionally robust

DROCS DR outpatient colonoscopy scheduling

i.i.d. independent and identically distributed

SAA sample average approximation

SASS single-server stochastic appointment sequencing and scheduling

SMILP stochastic mixed-integer linear program

SOASP stochastic outpatient appointment scheduling problem

SOPSP stochastic outpatient procedure scheduling problem

SP stochastic programming

TSM two-stage model

TSM-AO TSM under the appointment order

TSM-PI TSM under perfect information

MCO Monte-Carlo Optimization

MILP mixed-integer linear program

MSMIP multi-stage stochastic mixed integer program

MINLP mixed-integer nonlinear programming

NS Neighbor Swapping

OPC outpatient clinic

UM-MPU University of Michigan Medical Procedures Unit

ABSTRACT

Outpatient clinics (OPCs) are quickly growing as a central component of the healthcare system. OPCs offer a variety of medical services, with benefits such as avoiding inpatient hospitalization, improving patient safety, and reducing costs of care. However, they also introduce new challenges for appointment planning and scheduling, primarily due to the heterogeneity and variability in patient characteristics, multiple competing performance criteria, and the need to deliver care within a tight time window. Ignoring uncertainty, especially when designing appointment schedules, may have adverse outcomes such as patient delays and clinic overtime. Conversely, accounting for uncertainty when scheduling has the potential to create more efficient schedules that mitigate these adverse outcomes. However, many challenges arise when attempting to account for uncertainty in appointment scheduling problems. In this dissertation, we propose new stochastic optimization models and approaches to address some of these challenges.

Specifically, we study three stochastic outpatient scheduling problems with broader applications within and outside of healthcare and propose models and methods for solving them. We first consider the problem of sequencing a set of outpatient procedures for a single provider (where each procedure has a known type and a random duration that follows a known probability distribution), minimizing a weighted sum of waiting, idle time, and overtime. We elaborate on the challenges of solving this complex stochastic, combinatorial, and multi-criteria optimization problem and propose a new stochastic mixed-integer programming model that overcomes these challenges in contrast to the existing models in the literature. In doing so, we show the art of, and the practical need for, good mathematical formulations in solving real-world scheduling problems.

Second, we study a stochastic adaptive outpatient scheduling problem which incorporates the patients random arrival and service times. Finding a provably-optimal solution to this problem requires solving a multi-stage stochastic mixed integer program (MSMIP), which in turn must optimize a scheduling problem over each random arrival and service time for each stage. Given that this MSMIP is intractable, we present two approximation based on two-stage stochastic mixed-integer models and a Monte Carlo Optimization approach. In a series of numerical experiments, we demonstrate the near-optimality of the appointment order (AO) rescheduling policy, which requires that patients are served in the order of their scheduled appointments, in many parameter settings. We also identify parameter settings under which the AO policy is suboptimal. Accordingly, we propose an alternative swap-based policy that improves the solution of such instances.

Finally, we consider the outpatient colonoscopy scheduling problem, recognizing the impact of pre-procedure bowel preparation (prep) quality on the variability of colonoscopy duration. Data from a large OPC indicates that colonoscopy durations are bimodal, i.e., depending on the prep quality they can follow two different probability distributions, one for those with adequate prep and the other for those with inadequate prep. We define a distributionally robust outpatient colonoscopy scheduling (DRCOS) problem that seeks optimal appointment sequence and schedule to minimize the worst-case weighted expected sum of patient waiting, provider idling, and provider overtime, where the worst-case is taken over an ambiguity set characterized through the known mean and support of the prep quality and durations. We derive an equivalent mixed-integer linear programming formulation to solve DRCOS. Finally, we present a case study based on extensive numerical experiments in which we draw several managerial insights into colonoscopy scheduling.

CHAPTER 1

Introduction

An outpatient clinic (OPC) is a medical facility designed for the treatment of outpatients, patients who visit OPC for diagnostic or treatment purposes but do not at this time require hospital admission (i.e., overnight stay). OPCs are becoming a central component in the healthcare system in part because they offer benefits such as diverse surgical and non-surgical specialties (e.g., endoscopic procedures, chemotherapy treatment, etc.), avoiding inpatient hospitalization, high patient safety outcomes, and low costs of care (*Ahmadi-Javid et al.*, 2017).

Outpatient appointment planning and scheduling involve the design of a template schedule consisting of appointment slots, into which patients are later assigned, often on a first-request-first-scheduled basis (*Riise et al.*, 2016). Outpatient appointment scheduling problems are typically modeled as single server sequencing and scheduling problems, considering a single medical provider, and where a task or an activity represents each patient type. The quality of a schedule's performance is often a function of patient waiting time, provider idle time, and provider overtime.

Outpatient appointment scheduling problems are challenging, for a number of reasons. First, multiple (heterogeneous) patient types may require different service times. Even patients of the same type may vary in the required service time, and such variation is hard to predict in advance, which may contribute to patient delay, provider idling, and provider overtime (*Ahmadi-Javid et al.* (2017); *Alexopoulos et al.* (2008); *Cayirli and Yang* (2014); *Deceuninck et al.* (2018); *Glowacka et al.* (2017); *Gupta and Denton* (2008); *Klassen and Yoogalingam* (2014)).

Second, patients' lack of punctuality is a common phenomenon in OPC, and it introduces additional complexity into the design and analysis of outpatient scheduling systems and increases OPC operational costs (*Deceuninck et al.* (2018); *Glowacka et al.* (2017); *Klassen and Yoogalingam* (2014)). Very late arrivals often create provider idling (and thus poor utilization of the clinic resources) and, along with the random service durations, may increase the waiting time of the subsequent appointments and provider overtime through the propagation of the delay, if there is no buffer in the schedule to absorb it. Very early arrivals, on the other hand, may require the provider to make challenging queuing decisions whether to serve them ahead of their scheduled time (*Deceuninck et al.* (2018); *Glowacka et al.* (2017); *Samorani and Ganguly* (2016)).

Ignoring the heterogeneity and variability in patient characteristics when designing OPC appointment schedules may have adverse outcomes. For example, by only considering the average values of service durations we could schedule unnecessarily long (respectively, short) time in between appointments, resulting in significant provider idling and/or overtime (respectively, patient waiting). Conversely, accounting for uncertainty in the scheduling decision process has the potential to create efficient and robust schedules that mitigate these adverse outcomes. However, many challenges arise when attempting to model and solve appointment scheduling problems subject to uncertainty.

Despite the extensive amount of work that has been done within the field of stochastic (outpatient) appointment scheduling, challenges remain (*Ahmadi-Javid et al.*, 2017). This dissertation contains three main chapters, each of which focuses on different challenging "offline" stochastic outpatient scheduling problem with broader applications within and outside of healthcare ("offline" in the sense that we make all scheduling decisions ahead of time) and proposes new models and approaches for solving them. In the first two problems, the probability distributions of random parameters are known, and so we leverage the ideas and tools of stochastic programming (SP). In the SP approach, we look for scheduling decisions that minimize an expected weighted sum of the scheduling metrics, where the expectation is taken with respect to the "known" joint distribution of random parameters (see, e.g., *Birge and Louveaux* (2011); *Shapiro et al.* (2009) for a thorough introduction to SP).

In the third problem, the probability distributions of the random parameters are unknown. Therefore, we leverage the ideas and tools of distributionally robust (DR) optimization to address this uncertainty. In the DR approach, we look for scheduling decisions that minimize the worst-case expected weighted sum of the scheduling metrics. Here, the worst-case is taken over an ambiguity set. The ambiguity set is a family of distributions characterized by some known properties of the unknown probability distributions of uncertain parameters (see, e.g., *Bertsimas and Popescu* (2005); *Bertsimas et al.* (2010); *Ben-Tal and Nemirovski* (1998); *Delage and Ye* (2010); *Esfahani and Kuhn* (2018); *Scarf* (1958); *Shang and You* (2018) and references therein).

The remainder of this dissertation is structured as follows. In Chapter 2, we present a new stochastic mixed-integer linear program (SMILP) for the problem of sequencing a set of outpatient procedures for a single provider (where each procedure has a known type and a random duration that follows a known probability distribution associated with the procedure type) and determining the associated scheduled start time for each procedure. Our objective is to minimize the expectation of a weighted sum of patient waiting time, provider idling, and clinic overtime.

To provide context within the literature, we compare our SMILP model with those of *Berg et al.* (2014) (an enhancement of *Denton et al.* (2007)) and *Mancilla and Storer* (2012), which are, to the best of our knowledge, the only SMILPs for similar single-server stochastic appointment sequencing and scheduling (SASS) problems. We analyze the three models both theoretically and empirically, demonstrating where significant improvements in performance can be gained with our proposed model. In doing so, we show the art of, and the practical need for, good mathematical formulation in solving real-world scheduling (and mixed-integer programming) problems.

To the best of our knowledge, and according to the recent review of outpatient appointment systems by *Ahmadi-Javid et al.* (2017), Chapter 2 (*Shehadeh et al.*, 2019) presents the first rigorous and computational analysis of models for SASS with stochastic service duration.

One of the shortcomings of the scheduling models that we analyze in Chapter 2 is that they ignore the uncertainty pertaining to patients' arrival times and the possibility of rescheduling, i.e., resequencing to another position in the appointment sequence or declining to serve chronically late patients. Therefore, in Chapter 3, we study a more general stochastic outpatient appointment scheduling problem (SOASP) in which we incorporate the random patients' arrival times, random service durations, and adaptive rescheduling. Finding a provably optimal solution to this problem requires solving a multi-stage stochastic mixed integer program (MSMIP) with an initial schedule made in the first stage and rescheduling policy optimized in the subsequent stages.

In recognition that this MSMIP is intractable, we first consider a two-stage model (TSM) that relaxes the non-anticipativity constraints of MSMIP, thus yielding a lower bound. Second, we derive a family of valid inequalities to strengthen and improve the solvability of this TSM. Third, we obtain an upper bound for the MSMIP by solving another TSM model, under the feasible (and implementable) appointment order (AO) policy, which requires that patients are served in the order of their scheduled appointments, enforcing non-anticipativity. Fourth, we propose a Monte Carlo approach to evaluate the relative gap between these MSMIP bounds. Fifth, we show that the MSMIP bounds are very close in many SOASP parameter settings, demonstrating the near-optimality of the AO policy. We also identify parameter settings that result in a large gap. Accordingly, we close by proposing an alternative swap-based policy that improves the solution in such instances.

To the best of our knowledge, and according to the recent review of outpatient appointment systems by *Ahmadi-Javid et al.* (2017) and the literature review in Section 4.3, Chapter 3 presents the first stochastic programming approach to SOASP that considers (1) patient heterogeneity, (2) optimizing both the initial appointment sequencing and scheduling decisions, and (3) the possibility of rescheduling (i.e., resequencing or declining).

Finally, in Chapter 4, we consider the challenges of colonoscopy scheduling at the University of Michigan Medical Procedures Unit (UM-MPU), an OPC that performs a variety of endoscopic procedures including a large number of colonoscopies. Colonoscopy, in particular, is the mainstay of diagnosis and prevention for colorectal cancer (CRC), a leading cause of cancer-related deaths worldwide (*Anderson and Butterly*, 2015; *Singh et al.*, 2016; *Zauber et al.*, 2012).

Scheduling colonoscopy yields an added challenge and complexity to the problems we address in Chapters 2–3 for several reasons. First, there is significant variability in colonoscopy duration, primarily due to the quality of pre-procedure bowel preparation (prep) that the patient must undergo (*Bechtold et al.* (2016); *Chokshi et al.* (2012); *Froehlich et al.* (2005); *Johnson et al.* (2014); *Lebwohl et al.* (2010); *Rex et al.* (2002, 2006)). Our analysis of the UM-MPU data suggests that colonoscopy durations are "*bimodal*," i.e., depending on the prep quality, they can follow one of two different probability distributions, one for those with adequate prep and the other for those with inadequate prep (see this analysis in Section 4.2). Unfortunately, when scheduling a patient, it is not known at that time whether the patient will perform an adequate prep or not. Furthermore, there is a wide range of possible probability distributions for modeling the variability in colonoscopy duration with adequate and inadequate prep.

Second, colonoscopy is often scheduled with an upper endoscopy. The variability in the duration of the combined colonoscopy and upper endoscopy is primarily due to the variability in colonoscopy duration (as a function of bowel prep). Moreover, the duration of a combined procedure is longer than that of a colonoscopy procedure. This requires the OPC managers to make complex sequencing decisions about the order of colonoscopies and the combined upper endoscopy and colonoscopy procedures.

Third, colonoscopy outcome is a function of the time of the day, possibly as a consequence of provider fatigue as the day progress (see, e.g., *Almadi et al.* (2015); *Singh et al.* (2016)). As a result, the provider often has a preference for earlier start times for those patients who are at high risk of CRC. Accommodating provider preference and maintaining good operational performance are challenging to trade off. For example, scheduling the combined procedure of a high-risk patient first in the day may delay the start time of subsequent scheduled appointments.

Finally, the UM-MPU data shows significant variability in patient's actual arrival time relative to their scheduled arrival time and the distribution of arrival time deviations is unknown.

The bimodality and ambiguity in the distribution of colonoscopy duration (as a function of uncertain prep quality) prevent us from using the SP approaches in Chapters 2–3 (which assume that we know the probability distributions of uncertain parameters) for colonoscopy scheduling. We therefore define a DR outpatient colonoscopy scheduling (DROCS) problem that seeks optimal appointment sequence and schedule to minimize the worst-case expected weighted sum of patient waiting, provider idling, and provider overtime. Here, we take the worst-case over an ambiguity set characterized by the known mean and support of the prep quality and durations. We derive an equivalent mixed-integer linear program (MILP) formulation to solve DROCS.

Using the UM-MPU data, we then conduct extensive numerical experiments to draw insights into colonoscopy scheduling. Specifically, we demonstrate that this DR approach can produce schedules that (1) have a good operational performance (in terms of waiting time, idle time, and overtime) under various probability distributions (and extreme scenarios) of the random parameters, and (2) can accommodate provider (and patient) preference on appointment time while maintaining a good operational performance as compared to the SP approach.

To the best of our knowledge, and according to the recent review of outpatient appointment systems by *Ahmadi-Javid et al.* (2017), the work in Chapter 4 is the first to address the bimodal ambiguity of colonoscopy (service) durations. We further contribute with a new DR model that incorporates sequencing decisions and considers the ambiguity of two coexisting uncertainties of colonoscopy duration (as a function of uncertain prep quality) and arrival time deviation.

Collectively, this dissertation addresses four salient challenges to efficient outpatient appointment scheduling under uncertainty: random service duration (Chapter 2), random arrival time (Chapter 3), the possibility of rescheduling (Chapter 3), and bimodality and ambiguity of the distribution of service duration (Chapter 4). More broadly, this dissertation contributes to the literature on scheduling under uncertainty, and stochastic optimization with guidelines and methods to develop tractable and implementable scheduling (and mixed-integer programming) models and approaches for real-world optimization problems under uncertainty.

CHAPTER 2

Analysis of Models for the Stochastic Outpatient Procedure Scheduling Problem

2.1 Introduction

In this chapter, we address the stochastic outpatient procedure scheduling problem (SOPSP), which arises in outpatient procedure centers (OPCs). In this problem, we consider the perspective of an OPC manager who must schedule the start times for a day's worth of procedures for a single provider, where each procedure has a known type and a random (non-negative) duration that follows a known probability distribution associated with the procedure type. Given the uncertainty in procedure durations, the goal is to minimize the expectation of a weighted sum of total patient waiting time (the time from the scheduled start of a procedure to its actual start), total provider idle time (the time from the end of one procedure to the start of the next), and clinic overtime (the time from the scheduled closing time of the clinic to the end of the last procedure of the day).

The SOPSP is computationally challenging to solve, for a number of reasons. First, it is a complex combinatorial optimization problem, given the inherent implied sequencing problem that underlies assigning appointment times to each patient (*Ahmadi-Javid et al.*, 2017; *Berg et al.*, 2014; *Mancilla and Storer*, 2012). Second, the problem is inherently stochastic due to the uncertainty in procedure durations. Finally, it is also a multi-criteria optimization problem, in which we must make trade-offs between longer spacing between appointments, which leads to reduced patient delays, and shorter spacing, which leads to less provider idling and overtime (see, e.g., *Antunes et al.* (2016); *Marler and Arora* (2004); *T'kindt and Billaut* (2006) for a thorough discussion multi-criteria optimization). More broadly, the SOPSP is a single-server stochastic appointment sequencing and scheduling (SASS) problem, the underlying complexity of which has been studied by several previous authors beginning with the seminal work of *Welch and Bailey* (1952) and *Weiss* (1990) (see *Ahmadi-Javid et al.*, 2017; *Berg et al.*, 2014; *Denton et al.*, 2010; *Gupta*, 2007; *Gupta and Denton*, 2008; *Mancilla and Storer*, 2012, and references therein).

In addition to the value that the ability to solve this challenging SOPSP provides to OPC managers, it also has relevance for many other applications, including scheduling of surgeries in an operating room, ships in a port, exams in an examination facility, and more (*Ahmadi-Javid et al.*, 2017; *Begen and Queyranne*, 2011; *Mancilla and Storer*, 2012; *Robinson and Chen*, 2003; *Sabria and Daganzo*, 1989). For example, it is a common practice for surgeries to initially be assigned to a surgeon, date, and operating room several weeks or even months before their scheduled date. The actual scheduled start times for these surgeries, however, are typically not set until a few days in advance. It is at this point when the SOPSP can be solved to construct the final surgical schedule and notify the patients when to report to the hospital (see *Denton et al.*, 2010; *Mancilla and Storer*, 2012, and references therein for more details).

In this chapter, we present a new stochastic mixed-integer linear program (SMILP) using sample average approximation (SAA) for solving the SOPSP, with a focus both on *tractability* (i.e., being able to solve problem instances of realistic sizes in an acceptable amount of time) and *implementability* (i.e., proposing a model that can be easily translated into standard optimization software packages, not requiring customized algorithmic development or tuning). To provide context within the literature, we compare our model with those of *Berg et al.* (2014) (an enhancement of *Denton et al.*, 2007) and *Mancilla and Storer* (2012), which are, to the best of our knowledge, the only SMILPs for SASS with waiting, idling, and overtime costs. We discuss the relative strengths and weaknesses of the three models and then compare them computationally under a common, straightforward software implementation.

The remainder of this chapter is structured as follows. In Section 3.2, we present the relevant literature. In Section 2.3, we introduce and analyze three mathematical models of the SOPSP: two based on prior literature (*Berg et al.*, 2014 and *Mancilla and Storer*, 2012), and a new model. After that, in Section 2.4, we compare the computational performance of the three models and provide some discussion and insights. Finally, we conclude and summarise this chapter in Section 2.5.

2.2 Literature Review

Outpatient scheduling problems have been an active area of research since the seminal work of *Welch and Bailey* (1952). Comprehensive surveys of results obtained since then include *Cayirli and Veral* (2003), *Gupta and Denton* (2008), and *Ahmadi-Javid et al.* (2017). Within this literature, there are two primary approaches to stochastic appointment scheduling. The first is to develop and evaluate scheduling heuristics, often through the use of simulation (see, for example, *Ahmadi-Javid et al.*, 2017; *Ho and Lau*, 1992; *Klassen and Rohleder*, 1996; *Rohleder and Klassen*, 2000; *Vissers and Wijngaard*, 1979). The second is to construct models and design algorithms to find optimal schedules through the use of queueing theory (see, for example, *Bosch and Dietz*, 2000; *Jansson*, 1966; *Mercer*, 1960; *Sabria and Daganzo*, 1989; *Soriano*, 1966; *Vanden Bosch and Dietz*, 2001, and references therein), stochastic programming (see, for example, *Berg et al.*, 2014; *Denton and Gupta*, 2003; *Mancilla and Storer*, 2012; *Robinson and Chen*, 2003, and references therein), and, more recently, robust and distributionally robust optimization (RO and DRO, respectively; see, for example, *Jiang et al.*, 2017; *Mak et al.*, 2014, and references therein).

Herein, we present studies that are most relevant to this chapter: papers that use SMILP models to address offline single-resource stochastic appointment sequencing and scheduling (SASS) problems that are similar to the SOPSP ("offline" in the sense that sequencing and scheduling decisions are all made ahead of time). We are interested in generating optimal solutions to the SOPSP assuming knowledge of the distributions of appointment durations (a classic SASS assumption, *Ahmadi-Javid et al.*, 2017; *Berg et al.*, 2014; *Deceuninck et al.*, 2018), which rules out both the heuristic approach (due to sub-optimality and lack of performance guarantees, *Ahmadi-Javid et al.*, 2017; *Ho and Lau*, 1992; *Rohleder and Klassen*, 2000; *Klassen and Rohleder*, 1996; *Vissers and Wijngaard*, 1979) and the RO and DRO-based approaches (which assume distributional ambiguity). Finally, as pointed out by *Robinson and Chen* (2003), queueing theory-based results and algorithms are not appropriate for the SOPSP and other OPC scheduling problems which involve serving a finite number of patients within fixed service hours (i.e., the queue never reaches a steady state).

Papers that present models and algorithms for optimizing SASS decisions using SMILP fall into two groups: those that focus on determining the optimal start times (or, equivalently, the interarrival times) assuming that the sequence of patients (customers) is already fixed (e.g., through the use of a heuristic, see, for example, *Bosch and Dietz*, 2000; *Denton and Gupta*, 2003; *Erdogan and Denton*, 2013; *Ge et al.*, 2013; *Robinson and Chen*, 2003; *Vanden Bosch and Dietz*, 2001, and references therein), and those that focus on optimizing the sequencing and scheduling decisions simultaneously. Since we consider both sets of decisions, we further limit the scope of this review to the latter category. We refer the reader to the following studies: *Ahmadi-Javid et al.* (2017); *Berg et al.* (2014); *Cayirli et al.* (2006, 2008); *Creemers et al.* (2012); *Gupta and Denton* (2008); *Rohleder and Klassen* (2000); *Salzarulo et al.* (2016), and references therein, which demonstrate the benefit of sequencing heterogeneous patient appointments based on their characteristics for improving clinic performance and reducing costs compared to fixed sequence approaches. To the best of our knowledge, and according to the recent review of outpatient appointment systems by *Ahmadi-Javid et al.* (2017), papers by *Denton et al.* (2007), *Berg et al.* (2014), and *Mancilla and Storer* (2012) are the ones most closely related to our work, addressing similar SASS problems with waiting, idling, and overtime costs using SMILP.

Denton et al. (2007) formulated the stochastic surgery scheduling problem in an operating room (OR) as a two-stage SMILP with binary precedence variables and continuous time allowance variables in the first stage, and continuous waiting, idling, and overtime variables in the second stage. They used the sample-average approximation approach (i.e., a scenario-based approach) to replace the continuous distributions of surgery durations with approximate discrete distributions by considering a sample of N randomly generated scenarios. Since it was difficult to solve instances with more than 4 surgeries, they proposed several sequencing heuristics and then obtained the optimal surgery start times, for a fixed sequence, via the L-shaped algorithm (*Birge and Louveaux*, 2011) described in *Denton and Gupta* (2003). Their results showed substantial potential reductions in surgeon waiting, OR idling, and overtime costs by sequencing surgeries based on variances of their durations compared to the schedule of the OR that the study considered.

In a slightly different setting, *Berg et al.* (2014) considered the problem of optimizing the booking (number of patients to schedule) and appointment time decisions for outpatient procedures under no-show and procedure durations uncertainties. The goal was to maximize profit, i.e., the difference between the expected revenue and the expected variable cost of patient waiting time, provider idle time, and overtime associated with scheduling patients. Since the revenue was straightforward to compute, the paper focused on minimizing the expected variable cost determined by sequencing and scheduling decisions (a SASS problem which is, to some extent, similar to the SOPSP). To that end, the paper extended and enhanced the SMILP model of *Denton et al.* (2007) by including heterogeneous no-show probabilities and using both precedence and assignment variables to strengthen the earlier model, and employed three exact solution methods: L-shaped, hybrid multi-cut L-shaped with scenario aggregation and ranking (to overcome the computational burden of the original multi-cut method, see *Birge and Louveaux*, 1988), and branch-and-bound with progressive hedging as a primal heuristic (*Rockafellar and Wets*, 1991). While these methods were computationally competitive (relative to each other) in solving small instances ($\leq$ 5 patients), it was challenging to solve larger instances (10 patients), primarily due to the stochastic and combinatorial elements of the problem. Therefore, they proposed six sequencing heuristics based on standard deviations of procedure durations and no-show probabilities, and illustrated the conditions under which some of these provided a near-optimal solution to the problem.

Mancilla and Storer (2012) formulated the surgery sequencing and scheduling problem in a single operating room at a local hospital as a stochastic mixed-integer program with sample average approximation. The model differs from that of *Denton et al.* (2007) in the following two ways. First, they replaced binary precedence variables with binary sequence position assignment variables (previously proposed in *Wagner*, 1959). Second, they replaced continuous job time allowance variables with continuous appointment (start) time variables. Additionally, using concepts from *Garey et al.* (1976), they proved that for two scenarios and equal idling costs but different waiting costs for each job, the finite scenario SAA problem is NP-complete. Therefore, to overcome the computational burden of the sequencing decisions, they developed an algorithm to generate a near-optimal sequence, with the resulting linear subproblem of determining appointment times solved within their algorithm using the CPLEX barrier method. Given that the SMILP studied in *Mancilla and Storer* (2012) is a variation of the one in *Denton et al.* (2007), and the one in *Berg et al.* (2014) is stronger than *Denton et al.* (2007), in this chapter, we focus our analysis on the models of *Mancilla and Storer* (2012) and *Berg et al.* (2014).

Finally, we point out the similarities and differences between single provider stochastic appointment sequencing and scheduling and single machine scheduling (SMS). At the outset, they look similar: the provider can be thought of as a single machine, and procedures and their durations as jobs and their processing times, respectively (see *Forst*, 1993; *Lawler et al.*, 1993; *Pinedo*, 2016 for machine scheduling literature). Nevertheless, SASS is materially different from SMS. In SMS problems, each job release time (the time at which the job becomes available for processing) is typically exogenous (i.e., a parameter). In contrast, the appointment time in SASS, which can be thought of as a release time at which the scheduled patient is presumably available for the procedure, is a decision variable. Furthermore, in the classic SMS problem, one scheduling criterion that has received the most attention over the years is minimizing makespan (i.e., completing the last job at the earliest possible time), which trivially minimizes overtime but does not consider patient waiting time nor provider idle time. Our SMILP model, as well as those of *Mancilla and Storer* (2012) and *Berg et al.* (2014), however, improve on some ideas from the seminal work of *Wagner* (1959) and *Pinto and Grossmann* (1998) in the domain of deterministic single-machine jobs/tasks sequencing and scheduling.

2.3 Stochastic Mixed-Integer Linear Programming Models of the SOPSP

In this section, we present and analyze three SMILP formulations for the SOPSP. First, we define the problem formally. Then, we present our SMILP formulation and the conditions under which it is equivalent to two closely-related stochastic appointment sequencing and scheduling SMILPs in the literature, those of *Mancilla and Storer* (2012) and *Berg et al.* (2014), which are also presented for completeness.

2.3.1 Formal Statement of the Problem

We consider the problem of sequencing a set of procedures for a single provider (where each procedure has a known type and a random, non-negative, duration that follows a known probability distribution associated with the procedure type) and determining the associated scheduled start time for each procedure. The performance metric is the weighted sum of three components, total patient waiting time (the time from the scheduled start of a procedure to its actual start), total provider idle time (the time from the end of one procedure to the start of the next), and overtime (the time from the scheduled closing time of the clinic to the end of the last procedure of the day). Given a set of procedures, their sequence, their scheduled start times, and the distributions of their durations, the expected value of this weighted sum can be estimated by averaging over finitely many realizations (a sample) of procedure durations. This sample average is the objective function of the forthcoming optimization problems. We make the following assumptions:

A1 A procedure is not permitted to start before its scheduled start time nor the completion time of the previous procedure.

A2 Although patients may fail to show up to their appointments, we assume that those who do show up are punctual, i.e., available at the scheduled start times of their procedures.

A3 The provider is always available at the start of the day, and immediately after each procedure.

A4 There is no opportunity to modify the schedule on the day of service, i.e., rescheduling during the day or adding procedures (to accommodate walk-ins or emergencies) is not permitted.

The problem can be formulated as a two-stage SMILP with binary (for *sequencing*) and continuous (for *scheduling*, i.e., start times) first-stage variables and continuous second-stage variables representing what happens for each realization of procedure durations (waiting time, idle time, and overtime), given the sequence of appointment times decided in the first stage. To incorporate procedure duration uncertainty into the model, we use a sample average approximation (SAA) approach as in *Robinson and Chen* (2003), *Denton et al.* (2010), and *Mancilla and Storer* (2012). That is, we generate a sample of N scenarios (each scenario consists of a vector of realizations of procedure durations which are drawn independently from the distributions corresponding to each patient's type; a no-show patient can be represented by a realized procedure duration of 0), and then optimize the sample average of the weighted sum of the three metrics using the stored sample. (The technical details of sample average approximation approach are out of the scope of this chapter, and we refer the reader to *Homem-de Mello and Bayraksan*, 2014; *Kim et al.*, 2015; *Kleywegt et al.*, 2002; *Mak et al.*, 1999, and references therein, for a thorough discussion.)

Table 2.1: Notation Shehadeh et al. (2019).

Indices	
p	index of patients, or procedures, to be scheduled, $p = 1, \dots, P$
i	index of positions in the sequence, or appointments, $i = 1, \dots, P$
n	index of scenarios to be considered, $n = 1, \dots, N$
Parameters	
λ_i^w	waiting time penalty for appointment i
λ_i^g	penalty for idle time between appointments i and $i+1$
λ^o	overtime penalty
$\mathcal{L}$	planned length of clinic day
d_p^n	duration of procedure p in scenario n
Scenario-independent (first-stage) variables	
$x_{i,p}$	binary assignment variable indicating whether procedure p is assigned to appointment i
t_i	scheduled start time of appointment i
Scenario-dependent (second-stage) variables	
s_i^n	actual start time of appointment i in scenario n
g_i^n	idle time after appointment i in scenario n
o^n	overtime in scenario n.

2.3.2 Formulations of the Problem

Table 2.1 summarizes notation and some terminology used in our sample-average SMILP formulation of the SOPSP. Note, in particular, that we use the term “appointment” to refer to a position in the sequence, and use the terms “patient” and “procedure” interchangeably. Using this notation, the problem can be formulated as follows:

$$\text{(S) minimize } \frac{1}{N}\sum_{n=1}^{N}\left[\sum_{i=1}^{P}\lambda_i^w \cdot (s_i^n - t_i) + \sum_{i=1}^{P}\lambda_i^g \cdot g_i^n + \lambda^o \cdot o^n\right] \tag{2.1a}$$

$$\text{subject to } \sum_{i=1}^{P} x_{i,p} = 1 \qquad \forall p \tag{2.1b}$$

$$\sum_{p=1}^{P} x_{i,p} = 1 \qquad \forall i \tag{2.1c}$$

$$s_i^n \geq t_i \qquad \forall i, n \tag{2.1d}$$

$$s_i^n \geq s_{i-1}^n + \sum_{p=1}^{P} d_p^n \cdot x_{i-1,p} \qquad \forall (i \geq 2, n) \tag{2.1e}$$

$$g_i^n = s_{i+1}^n - \left(s_i^n + \sum_{p=1}^{P} d_p^n \cdot x_{i,p} \right) \qquad \forall (i < P, n) \tag{2.1f}$$

$$o^n \geq \left(s_P^n + \sum_{p=1}^{P} d_p^n \cdot x_{P,p} \right) - \mathcal{L} \qquad \forall n \tag{2.1g}$$

$$(g_i^n, s_i^n) \geq 0 \qquad \forall (i, n) \tag{2.1h}$$

$$o^n \geq 0 \qquad \forall n \tag{2.1i}$$

$$t_i \geq 0 \qquad \forall i \tag{2.1j}$$

$$x_{i,p} \in \{0, 1\} \qquad \forall (i, p) \tag{2.1k}$$

In the above formulation, the objective function in (2.1a) is the sample average of the weighted linear combination of the total waiting time, total idle time, and overtime cost. Constraints (2.1b) and (2.1c) ensure that each procedure is assigned to one appointment and each appointment is assigned one procedure. For every scenario n, constraints (2.1d) and (2.1e) require the actual start time, s_i^n, of the ith appointment to be no smaller than the scheduled start time, t_i, and than the completion time of the preceding appointment, i.e., the $(i-1)$st appointment's actual start time, s_{i-1}^n, plus the duration of the procedure assigned to it, $\sum_{p=1}^{P} d_p^n \cdot x_{i-1,p}$. The ith appointment waiting time is the difference between its actual and scheduled start time (i.e., $s_i^n - t_i$), which we include in the objective function directly. Constraints (2.1f) define the idle time between two consecutive appointments as the gap between the actual start time of an appointment and the completion time of the preceding one. Constraints (2.1g) and (2.1i) define overtime (if any) as the positive difference between the completion time of the last appointment and the clinic scheduled closing time, $\mathcal{L}$. Finally, the remaining constraints specify feasible ranges of the decision variables.

The formulation of *Mancilla and Storer* (2012) uses additional notation presented in Table 2.2. Note that components of g are indexed differently in this model than in our formulation (2.1a)–

Table 2.2: Additional notation Mancilla and Storer (2012).

Parameters	
λ_p^w	waiting time penalty for procedure p
λ_p^g	idle time penalty for procedure p
Scenario-dependent (second-stage) variables	
$w_{i,p}^n$	waiting time of procedure p in scenario n, if it is assigned to appointment i (0 otherwise)
$g_{i,p}^n$	idle time after procedure p in scenario n, if it is assigned to appointment i (0 otherwise)
e^n	slack variable measuring early completion of the schedule in scenario n

(2.1k), but this slight abuse of notation allows us to emphasize the relationship between two sets of variables representing idling times in the two models. The formulation of *Mancilla and Storer* (2012) is as follows:

$$\text{(M) minimize } \frac{1}{N}\sum_{n=1}^{N}\left[\sum_{i=1}^{P}\sum_{p=1}^{P}\lambda_p^w\cdot w_{i,p}^n+\sum_{i=1}^{P}\sum_{p=1}^{P}\lambda_p^g\cdot g_{i,p}^n+\lambda^o\cdot o^n\right] \tag{2.2a}$$

$$\text{subject to } \sum_{i=1}^{P}x_{i,p}=1 \qquad \forall p \tag{2.2b}$$

$$\sum_{p=1}^{P}x_{i,p}=1 \qquad \forall i \tag{2.2c}$$

$$t_i-t_{i+1}-\sum_{p=1}^{P}w_{i+1,p}^n+\sum_{p=1}^{P}g_{i,p}^n+\sum_{p=1}^{P}w_{i,p}^n=-\sum_{p=1}^{P}d_p^n\cdot x_{i,p} \qquad \forall(i<P,n) \tag{2.2d}$$

$$t_P+\sum_{p=1}^{P}w_{P,p}^n-o^n+e^n=-\sum_{p=1}^{P}d_p^n\cdot x_{P,p}+\mathcal{L} \qquad \forall n \tag{2.2e}$$

$$w_{i,p}^n\leq M_1^i\cdot x_{i,p} \qquad \forall(i,p,n) \tag{2.2f}$$

$$g_{i,p}^n\leq M_2\cdot x_{i,p} \qquad \forall(i,p,n) \tag{2.2g}$$

$$(w_{i,p}^n,g_{i,p}^n,o^n,e^n)\geq 0 \qquad \forall(i,p,n) \tag{2.2h}$$

$$t_i\geq 0 \qquad \forall i \tag{2.2i}$$

$$x_{i,p}\in\{0,1\} \qquad \forall(i,p) \tag{2.2j}$$

As described in *Mancilla and Storer* (2012), the objective function in (2.2a) is the sample average of the weighted linear combination of the total waiting cost, total idling cost, and overtime cost. Constraints (2.2b) and (2.2c) ensure that each procedure is assigned to one appointment, and each appointment is assigned one procedure. Constraints (2.2d) define, for each scenario, the waiting and idle time for every appointment. Constraints (2.2e) define overtime in scenario n. Constraints (2.2f) and (2.2g) are logical constraints that enforce the relationship between variables $w_{i,p}^n$, $g_{i,p}^n$, and $x_{i,p}$ (here, M_1^i, $i = 1, \ldots, P$, and M_2 are sufficiently large constants). Finally, the remaining constraints specify feasible ranges of the decision variables.

It is well known that, in order to strengthen the formulation, the values of "Big-M" constants in constrains such as (2.2f) and (2.2g) should be as small as possible without loss of optimality. *Mancilla and Storer* (2012) recommend setting

$$M_1^i = \sum_{j=1}^{i-1} \delta_j, \; i = 1, \ldots, P,$$

where δ_j corresponds to the jth largest value of $\max\limits_{n=1,\ldots,N} d_r^n - \min\limits_{n=1,\ldots,N} d_r^n$ over $r = 1, \ldots, P$, and

$$M_2 = \max_{p=1,\ldots,P} \left\{ \max_{n=1,\ldots,N} d_p^n - \min_{n=1,\ldots,N} d_p^n \right\}.$$

We followed this suggestion in our computational experiments in Section 2.4.

The formulation of *Berg et al.* (2014) uses additional notation defined in Table 2.3, and is as follows:

$$\text{(B) minimize } \frac{1}{N} \sum_{n=1}^{N} \left[\sum_{p=1}^{P+1} \sum_{p'=1}^{P} \lambda_{p,p'}^w \cdot A_{p'}^n w_{p,p'}^n + \sum_{p=1}^{P+1} \sum_{p'=1}^{P+1} \lambda_{p,p'}^g \cdot g_{p,p'}^n + \lambda^o \cdot o^n \right] \tag{2.3a}$$

$$\text{subject to } \sum_{p'=1}^{P+1} r_{p,p'} \leq 1 \qquad \forall p \tag{2.3b}$$

$$\sum_{p=1}^{P+1}\sum_{p'=1}^{P+1} r_{p,p'} = P \tag{2.3c}$$

$$x_{i,p} + x_{i+1,p'} - 1 \leq r_{p,p'} \qquad \forall (p, p', i \leq P) \tag{2.3d}$$

$$\sum_{i=1}^{P+1} x_{i,p} = 1 \qquad \forall p \tag{2.3e}$$

$$\sum_{p=1}^{P+1} x_{i,p} = 1 \qquad \forall i \tag{2.3f}$$

$$\sum_{p=1}^{P+1} r_{p,P+1} = 1 \tag{2.3g}$$

$$\sum_{p=1}^{P+1} r_{P+1,p} = 0 \tag{2.3h}$$

$$x_{P+1,P+1} = 1 \tag{2.3i}$$

$$w^n_{p,p'} \leq M^n_1 r_{p,p'} \qquad \forall (p, p', n) \tag{2.3j}$$

$$g^n_{p,p'} \leq M_2 r_{p,p'} \qquad \forall (p, p', n) \tag{2.3k}$$

$$-\sum_{p'=1}^{P+1} w^n_{p',p} + \sum_{p'=1}^{P+1} w^n_{p,p'} - \sum_{p'=1}^{P+1} g^n_{p,p'} = A^n_p d^n_p - y_p \qquad \forall (p : p \leq P, n) \tag{2.3l}$$

$$\sum_{p=1}^{P+1}\sum_{p'=1}^{P} g^n_{p,p'} - o^n + e^n = \mathcal{L} - \sum_{p=1}^{P+1} A^n_p d^n_p \qquad \forall n \tag{2.3m}$$

$$r_{p,p'}, x_{i,p} \in \{0, 1\}, y_p \geq 0 \qquad \forall (p, p', i) \tag{2.3n}$$

$$(w^n_{p,p'}, g^n_{p,p'}, o^n, e^n \geq 0) \qquad \forall (p, p', n) \tag{2.3o}$$

As described in Berg *et al.* (2014), this formulation uses a dummy procedure $P + 1$ that has zero duration and is always assigned to the appointment slot $P + 1$. The objective function in (2.3a)

Table 2.3: Additional notation Berg et al. (2014).

Indices	
p, p'	indices for procedures, $p, p' = 1, \dots, P+1$
i	index for appointments, $i = 1, \dots, P+1$
Parameters	
$\lambda^w_{p,p'}$	sequence-dependent waiting cost for procedure p' following procedure p
$\lambda^g_{p,p'}$	sequence-dependent cost of idling between procedures p and p'
A^n_p	binary attendance indicator for patient p in scenario n
Scenario-independent (first-stage) variables	
$r_{p,p'}$	binary precedence variable; equals 1 if and only if procedure p is followed by procedure p'
y_p	time allotted to procedure p
Scenario-dependent (second-stage) variables	
$w^n_{p,p'}$	sequence-dependent waiting time for procedure p' when preceded by procedure p in scenario n
$g^n_{p,p'}$	sequence-dependent idle time between procedures p and p' in scenario n
e^n	slack variable measuring early completion of the schedule in scenario n

is the sample average of the weighted linear combination of the total waiting cost, total idling cost, and overtime cost. Constraints (2.3b) ensure that each procedure precedes at most one other procedure. Constraints (2.3c) ensure that every procedure, except for the dummy procedure and the first procedure, is included in exactly two precedence relationships. Constraints (2.3d) state that a precedence relationship can only exist if that same relationship is defined by the appointment assignment decisions. Constraints (2.3e) and (2.3f) require that each procedure is assigned to one appointment, and each appointment is assigned one procedure. Constraints (2.3g)–(2.3i) ensure that the dummy procedure will be the last procedure as defined by the binary precedence variables and the appointment slot assignment variables. If procedure p does not precede procedure p', the associated sequence-dependent waiting and idle times will be 0 by constraints (2.3j) and (2.3k), where M^n_1 and M_2 are sufficiently large constants.

Constraints (2.3l) calculate the waiting and idle times associated with each procedure based on the waiting time for the preceding procedure. The clinic's overtime is defined by (2.3m). Finally, the remaining constraints specify feasible ranges of the decision variables. *Berg et al.* (2014) set $M_1^n = \sum_{p=1}^{P} d_p^n, n = 1, \ldots, N$, and M_2=$\mathcal{L}$, which we also used in our computation experiments in Section 2.4.

In the following discussion, we will refer to formulation (2.1) proposed in this chapter as (S) (for Shehadeh et al.), and to formulations (2.2) of *Mancilla and Storer* (2012) and (2.3) of *Berg et al.* (2014) as (M) and (B), respectively.

Note that each of the three models has different capabilities in handling various waiting and idling cost structures. Our model (S) can handle situations where the costs are appointment-specific, model (M) can handle situations where the costs are patient-specific, and model (B) can handle situations where the costs depend on the sequence of patients in the schedule.

We also note that the models take different approaches to calculating waiting times and costs in the presence of no-shows: both in model (M) and our model (S), waiting cost is incurred if an appointment runs late, even if the patient assigned to the following appointment does not show (indeed, a no-show patient is treated as a procedure with duration 0), while in model (B) no waiting cost is incurred in this situation.

In the remainder of this chapter, we will consider the SOPSP under the following additional assumptions: (i) zero no-show rate (i.e., $A_p^n = 1 \ \forall (p, n)$); (ii) identical waiting costs across appointments and procedures, i.e., $\lambda_i^w = \lambda^w \ \forall i, \ \lambda_p^w = \lambda^w \ \forall p,$ and $\lambda_{p,p'}^w = \lambda^w \ \forall (p, p')$; and (iii) identical idling costs across appointments and procedures, i.e., $\lambda_i^g = \lambda^g \ \forall i, \ \lambda_p^g = \lambda^g \ \forall p,$ and $\lambda_{p,p'}^g = \lambda^g \ \forall (p, p')$. Under these assumptions, models (S), (M), and (B) are SMILP formulations of the same SOPSP and are, therefore, equivalent. Table 2.4 presents the respective sizes, in terms of number of variables and constraints, of the three formulations under these assumptions.

Table 2.4: Sizes of formulations of the SOPSP with P procedures and N scenarios.

	(B)	**(M)**	**(S)**
# Binary variables	$2P^2 + 4P + 2$	P^2	P^2
# Continuous variables	$P + 1 + N(2P^2 + 4P + 4)$	$P + N(2P^2 + 2)$	$P + N(2P + 1)$
# First-stage constraints	$P^3 + 5P^2 + 11P + 10$	$P^2 + 3P$	$P^2 + 3P$
# Second-stage constraints	$N(4P^2 + 9P + 5)$	$N(4P^2 + P + 2)$	$5NP$

2.4 Computational Experiments

In this section, we present computational experiments that explore the size and characteristics of the SOPSP instances that can be solved with the three SMILP formulations presented in Section 2.3.2. In Section 2.4.1, we describe the set of the SOPSP instances that we constructed for our experiments, explain how we generated a testbed of sample average approximations (SAAs) for each instance, and discuss other experimental settings. We then present results in Section 2.4.2, comparing the computational performance of the three formulations.

2.4.1 Description of Experiments

To study the impact of a variety of problem characteristics on computational performance, we developed a set of divers SOPSP instances, in part based on prior literature, summarized in Table 2.5. Each of the 14 instances is characterized by the number of procedures to be scheduled, the types of procedures, and the number of procedures of each type (for example, Instance 1 involves scheduling 4 procedures: two of type A, one of type C, and one of type J). Probability distributions of procedure durations by type are contained in Table 2.6.

Instances 1–8, 10, and 11 were based on the data set provided as part of the AIMMS-MOPTA 5th Optimization Modeling Competition. For each procedure type, we used all procedure duration realizations provided in the data set to fit all valid parametric distributions using the open

Table 2.5: Characteristics of SOPSP instances.

Instance	**# of Procedures**	**# of Types**	**Procedures to be scheduled (by type)**
1	4 procedures	3 types	(2A, 1C, 1J)
2	5 procedures	4 types	(2A, 1G, 1H, 1J)
3	5 procedures	4 types	(1A, 1D, 2G, 1J)
4	6 procedures	5 types	(1A, 1B, 1F, 2G, 1H)
5	7 procedures	5 types	(1C, 1D, 1F, 1H, 3J)
6	7 procedures	6 types	(1A, 1B, 1D, 1E, 2G, 1J)
7	10 procedures	6 types	(3A, 1C, 1D, 1G, 1I, 3J)
8	10 procedures	6 types	(2A, 1B, 1D, 2G, 2I, 2J)
9	10 procedures	2 types	(6CL, 4U)
10	11 procedures	8 types	(2A, 1C, 2E, 1F, 1G, 1H, 2I, 1J)
11	11 procedures	6 types	(2A, 2F, 1G, 2H, 2I, 2J)
12	12 procedures	2 types	(9R, 3N)
13	16 procedures	2 types	(12R, 4N)
14	20 procedures	2 types	(15R, 5N)

source Matlab function `allfitdist` (*Sheppard*, 2012), selecting the distribution with the best combination of the reported Goodness of Fit metrics (e.g., Akaike Information Criterion, Bayesian Information Criterion, Negative of the Log Likelihood). Instance 9 was based on the problem studied by *Berg et al.* (2014), which includes procedures of two types: colonoscopies (CL) and upper endoscopies (U). Instances 12–14 were based on the problem studied in *Deceuninck et al.* (2018), where 75% of the patients are newly referred (N) and the remaining 25% are follow-up return (R) patients. Accordingly, we constructed instances with up to 20 procedures, since this is by far the maximum number of patients a single provider can see in a clinic session. In each instance, we set $\mathcal{L}$ equal to the expected total duration of the P procedures, as is done in *Mancilla and Storer* (2012), *Berg et al.* (2014), and others.

We considered three different sets of weights for the multi-criteria objective function: (i) $\lambda^w = \lambda^g = \lambda^o$; (ii) $\lambda^w = 1,\ \lambda^g = 0,\ \lambda^o = 10$; and (iii) $\lambda^w = 1,\ \lambda^g = 5,\ \lambda^o = 7.5$. For the first set of weights, each of the three objectives is equally important. The second set comes from *Berg et al.* (2014), where it was motivated by the argument that instances with $\lambda^g \neq 0$ proved to be computationally easier. The third set comes from *Deceuninck et al.* (2018), where the authors assumed that the overtime cost is 50% higher than the regular idling cost based on the OPC literature and

Table 2.6: Distribution information for procedure duration, by type.

Procedure type	Mean	Variance	Distribution
A	9.83	12.08	Lognormal
B	81.46	804.56	Normal
C	59.75	652.69	Lognormal
D	34.53	303.94	Lognormal
E	120.84	2.38e+3	Lognormal
F	47.76	232.06	Lognormal
G	43.94	469.86	Gamma
H	39.90	129.28	Lognormal
I	95.13	2.430e+3	Lognormal
J	19.51	99.36	Lognormal
U	12.05	188.57	Weibull
CL	30.96	58.75	Weibull
R	20.00	256.00	Lognormal
N	30.00	576.00	Lognormal

practice (*Cayirli et al.*, 2006; *Deceuninck et al.*, 2018). Note that, with these sets of weights, and assuming zero no-show rate, formulations (S), (M), and (B) are equivalent.

We added symmetry-breaking constraints (see *Denton et al.*, 2010; *Berg et al.*, 2014; *Ostrowski et al.*, 2011) to all three models, recognizing that the durations of procedures of the same type are identically distributed. In particular, let P_q be the set of procedures of type q, $q = 1, \dots, Q$. Without loss of generality, we can assume that procedures within each P_q are numbered sequentially. We added the following symmetry-breaking constraints to all three models:

$$x_{i,p} - \sum_{j>i}^{P} x_{j,p+1} \leq 0 \; \forall i = 1, ..., P, \; \forall p : p, p+1 \in P_q, \; q = 1, \; \dots, Q, \tag{2.4}$$

indicating that, if procedures p and $p+1$ are of the same type, p is scheduled before $p+1$.

For each of the 14 SOPSP instances and 3 sets of objective function weights, we generated 10 SAAs, for a total of 420 SAA instances, each with N =1,000 scenarios. Our choice of the sample size N was motivated by the trade-off between the computational effort required to solve the resulting mixed-integer linear programs (MILPs) and the quality of approximation of the expected value objective of SOPSP by its sample average. On the one hand, the sizes of MILP instances of (S), (M), and (B) increase with N (see Table 2.4), and their solution times increase as well. As demonstrated in Section 2.4.2, using formulation (S), we were able to solve all the SAAs associated with the SOPSP instances described in Table 2.5 with N =1,000 in a reasonable time.

On the other hand, optimal solutions of SAA instances with larger values of N are likely to be closer to optimality with respect to the expected value objective of SOPSP. The research literature on sample average approximation methods in stochastic optimization provides theoretical insights as well as guidance for selecting a sample size from this perspective. In particular, the so-called Monte Carlo Optimization (MCO) procedure can be used to calculate statistical lower and upper bounds on the optimal value of SOPSP based on an optimal solution to its SAA approximation, which in turn provide a statistical estimate of the relative approximation gap between the optimal value of SOPSP and its SAA approximation (see *Homem-de Mello and Bayraksan*, 2014 and *Kleywegt et al.*, 2002 and references therein for the description of the MCO methodology and other technical details.).

Applying the MCO procedure to the formulation (S) with N =1,000, we estimated the relative approximation gaps for the SOPSP instances described in Table 2.5 to range between 0.004% and 0.9%, whereas larger sample sizes resulted in longer solution times without consistent and significant improvements in the relative approximation gaps. Based on the above considerations, we selected N =1,000 for our computational experiments.

We represented and solved the 420 SAA instances using the AMPL modeling language and IBM ILOG CPLEX Optimization Studio (version 12.6.2). We used the default settings of the solver since our experiments showed no consistent benefits of any parameter or settings tuning. We imposed a solver time limit of 7,200 second (2 hours) for each SAA instance. We performed all experiments on an HP workstation running Windows Server 2012 with two 2.10GHz Intel E5-2620-v4 processors, each with 8 cores (16 total) and 128 GB shared RAM.

Table 2.7: Solution times (in seconds) using model (S)

	$\lambda^w = \lambda^g = \lambda^o$			$\lambda^w = 1, \lambda^g = 0, \lambda^o = 10$			$\lambda^w = 1, \lambda^g = 5, \lambda^o = 7.5$		
Inst	Min	Avg±stdv	Max	Min	Avg±stdv	Max	Min	Avg±stdv	Max
1	2	3±0.34	3	3	3±1	7	3	3±0.2	7
2	10	13±2	17	8	11±3	17	4	5±0.9	7
3	8	9±0.9	11	5	5±0.4	6	5	6±0.6	7
4	33	41±6	55	21	23±2	26	23	25±2	28
5	53	65±9	77	44	51±6	60	41	49±5	57
6	99	111±7	122	52	58±8	80	57	70±8	79
7	215	276±46	334	153	176±36	276	168	197±28	248
8	237	284±24	310	140	170±29	242	205	226±18	269
9	57	70±8	85	44	55±6	61	46	53±4	58
10	588	769±105	937	178	226±37	293	233	270±33	342
11	660	770±37	987	254	357±61	460	251	326±43	375
12	83	107±12	123	70	78±5	86	100	116±11	130
13	363	466±59	551	242	297±35	349	455	512±55	602
14	862	1218±164	1464	930	1189 ±193	1500	461	549 ± 76	703

2.4.2 Discussion of Results

Recall that formulation (2.1) proposed in this chapter is designated by (S), and formulations (2.2) of *Mancilla and Storer* (2012) and (2.3) of *Berg et al.* (2014) are designated by (M) and (B), respectively. Henceforth, we will assume that constraints (3.8) are included in each of the models.

Using our proposed model (S), we were able to solve all 420 instances of the SAAs associated with the SOPSP instances described in Table 2.5 within the imposed time limit of two hours. In fact, solution times of the SAAs that correspond to instances 1–9, 10–11 under the second and third weight sets, and 12– 13 were less than 10 minutes (see Table 2.7 for details). Moreover, solution times of the SAAs that correspond to the largest (in terms of the number of procedures) and the most complex SOPSP instance (which is somewhat less commonly encountered in practice), instance 14, were less than 25 minutes. These solution times are sufficient for real-world implementation of model (S). Below, we compare the computational performance of model (S) with models (M) and (B).

Table 2.8: Ratios of solution times of models (B) and (S) on SAAs solved by both.

$\lambda^w = \lambda^g = \lambda^o$ (a)			$\lambda^w = 1, \lambda^g = 0, \lambda^o = 10$ (b)			$\lambda^w = 1, \lambda^g = 5, \lambda^o = 7.5$ (b)		
Min	Avg±stdv	Max	Min	Avg±stdv	Max	Min	Avg±stdv	Max
6	31±29	116	4	33±27	107	8	51±35	138

[a] SOPSP Instances 1–6, 10 SAA instances each.
[b] SOPSP Instances 1–5, 10 SAA instances each.

Table 2.9: Ratios of solution times of models (B) and (S) on SAAs solved by both.

$\lambda^w = \lambda^g = \lambda^o$ (a)			$\lambda^w = 1, \lambda^g = 0, \lambda^o = 10$ (b)			$\lambda^w = 1, \lambda^g = 5, \lambda^o = 7.5$ (b)		
Min	Avg±stdv	Max	Min	Avg±stdv	Max	Min	Avg±stdv	Max
6	31±29	116	4	33±27	107	8	51±35	138

[a] SOPSP Instances 1–6, 10 SAA instances each.
[b] SOPSP Instances 1–5, 10 SAA instances each.

2.4.2.1 Comparison with Model (B) of *Berg et al.* (2014)

Using model (B), we were able to solve 160 of the 420 SAA instances to optimality within two hours, namely, all 60 SAAs that correspond to SOPSP Instances 1–6 and the first weight set, and all 100 SAAs that correspond to Instances 1–5 with the second and third weight sets. We present a comparison of solution times of these 160 SAAs by models (S) and (B) in Table 2.9. Observe that model (B) takes from 6 to 138 times longer than model (S). We attribute the difference in solution times to two primary reasons. First, as shown in Table 2.4, model (B) has significantly more variables and constraints. As argued by *Artigues et al.* (2015); *Catanzaro et al.* (2015); *Fortz et al.* (2017); *Jünger et al.* (2009); *Keha et al.* (2009); *Klotz and Newman* (2013); *Morales-España et al.* (2016); *Pochet and Wolsey* (2006), this increase in model size often suggests an increase in solution time for the linear programming (LP) relaxations. Second, as shown in Table 2.10, for all 420 SAAs, the LP relaxations obtained using model (S) were strictly tighter than using model (B), by a factor of 1.11 to 3.48.

Finally, for the 260 SAAs that were not solved by model (B) in two hours, we report the relative MIP (relMIP) gap, calculated as relMIP gap $= \frac{\text{UB}-\text{LB}}{\text{UB}} \times 100\%$, where UB is the best upper bound and LB is the linear programming relaxation-based lower bound obtained at termination after 2

Table 2.10: Ratios of optimal objective values of LP relaxations of (S) and (B).

$\lambda^w = \lambda^g = \lambda^o$			$\lambda^w = 1, \lambda^g = 0, \lambda^o = 10$			$\lambda^w = 1, \lambda^g = 5, \lambda^o = 7.5$		
Min	Avg±stdv	Max	Min	Avg±stdv	Max	Min	Avg±stdv	Max
1.95	2.62±0.41	3.48	1.11	1.38±0.26	2.08	1.27	1.64±0.33	2.49

Table 2.11: Relative MIP gap at termination for SAAs not solved by (B) in two hours.

$\lambda^w = \lambda^g = \lambda^o$(a)			$\lambda^w = 1, \lambda^g = 0, \lambda^o = 10$(b)			$\lambda^w = 1, \lambda^g = 5, \lambda^o = 7.5$(b)		
Min	Avg±stdv	Max	Min	Avg±stdv	Max	Min	Avg±stdv	Max
41%	54±0.08%	70%	19%	34±0.09%	53%	16%	40±0.09%	52%

[a] SOPSP Instances 7–12, 10 SAA instances each.
[b] SOPSP Instances 6–11, 10 SAA instances each.

hours. Of the 260 SAAs in question, 180 terminated with a relMIP gap between 16% and 70% (see Table 2.11 for details), while the remaining 80 SAAs terminated without any feasible MIP solutions (and thus no upper bound).

2.4.2.2 Comparison with Model (M) of *Mancilla and Storer* (2012)

Using model (M), we solved 340 of the 420 SAAs to optimality within the two hour time limit. We present performance comparisons for these instances in Table 2.12. Table 2.13 identifies the SOPSP instances that gave rise to the remaining 80 SAAs.

In exploring the difference in solution times between the two models, we first observe that they have the same first-stage formulation. Furthermore, as we prove in Theorem 2.6.1 in Appendix 2.6.1, the LP relaxations of the two models have the same optimal objective values. In fact, using the same proof techniques, we can show that, given any set of values of variables $x_{i,p}\ \forall(i,p)$ that satisfy constraints (2.1b) and (2.1c) (which are identical to constraints (2.2b) and (2.2c)) and $0 \leq x_{i,p} \leq 1\ \forall(i,p)$, the optimal objective value obtained by optimizing the remaining (continuous) variables will be the same for either model. This suggests that a branch-and-bound algorithm would perform similarly on both models in terms of the number of nodes explored (recognizing that there will be variability due to CPLEX preprocessing and implementation of branch-and-cut

Table 2.12: Comparison of performance of models (M) and (S) on SAAs solved by both: solution time, number of nodes, simplex iterations.

	$\lambda^w = \lambda^g = \lambda^o$			$\lambda^w = 1, \lambda^g = 0, \lambda^o = 10$			$\lambda^w = 1, \lambda^g = 5, \lambda^o = 7.5$		
Ratio	Min	Avg±stdv	Max	Min	Avg±stdv	Max	Min	Avg±stdv	Max
$\frac{\text{(M) sol. time}}{\text{(S) sol. time}}$	1.2	7±4	21	2	13±9	43	1.1	7±5	27
$\frac{\text{(M) nodes}}{\text{(S) nodes}}$	0.5	1±0.2	1.4	0.2	1±0.3	1.9	0.4	1±0.2	1.4
$\frac{\text{(M) iterations}}{\text{(S) iterations}}$	1	11±15	119	1	12±19	133	1	16±22	113

Table 2.13: Number of SAA instances that were not solved to optimality in the two hours by model (M).

SOPSP Instance #	$\lambda^w = \lambda^g = \lambda^o$	$\lambda^w = 1, \lambda^g = 0, \lambda^o = 10$	$\lambda^w = 1, \lambda^g = 5, \lambda^o = 7.5$
10	10	5	4
11	10	10	0
13	6	2	3
14	10	10	10

instead of a traditional branch-and-bound). The ratios between the number of nodes explored by CPLEX for the two models for the 340 SAAs solved by both are, indeed, on average equal to 1 for each of the weight sets, as reported in Table 2.12.

Clearly, then, the difference in solution times between models (S) and (M) is primarily due to differences in time spent exploring each node. This is supported further by Table 2.12 which reports the ratios in the numbers of simplex iterations required to solve each instance using the two models. The number of iterations is typically much larger for model (M), presumably as a result of the significantly larger second-stage formulation (see Table 2.4).

Finally, for the 80 SAAs that were not solved by model (M) in 2 hours, the relMIP gap at termination was 15% on average, with the maximum of 25%.

2.5 Conclusion and Chapter Summary

In this chapter, we presented a new stochastic mixed-integer linear programming model for the SOPSP using a sample-average approximation. This problem considers the perspective of an OPC manager who must schedule the start times for a day's worth of procedures (patients) for a single provider, where each procedure has a known type and a random (non-negative) duration that follows a known probability distribution associated with the procedure type. Given the uncertainty in procedure duration, the goal is to minimize the expectation of a weighted sum of patient waiting time, provider idle time, and clinic overtime. Our model allows for appointment-dependent waiting and idling costs, and treats patient no-shows as procedures with duration 0.

The SOPSP is a basic (yet still challenging) offline single-resource stochastic sequencing and scheduling problem that has been studied in various forms by several previous authors. Therefore, we compared our model with two closely-related models by *Mancilla and Storer* (2012) and *Berg et al.* (2014) under assumptions that ensure their equivalence, and analyzed them both empirically and theoretically. Computational results demonstrated where significant improvements in performance could be gained with our proposed model.

In addition to empirical tractability, our modeling approach has the advantage of implementability. Indeed, our proposed model performed well in the computational experiments that were performed using commonly available computer resources, a standard optimization modeling tool, and a commercial MILP solver with default settings — in other words, it did not require development of any specialized algorithms or a time-consuming search for beneficial software parameter settings. This is in contrast to previously-studied models of *Mancilla and Storer* (2012) and *Berg et al.* (2014), which were used in conjunction with specially-developed algorithms or heuristics in the original papers, but did not perform as well as our model with straightforward implementation. Implementability in the above sense is necessary for an optimization-based decision support tool to gain wide adoption in OPCs and other healthcare systems that do not have ongoing access to support staff with optimization expertise, and thus is a valuable feature of our proposed model.

To the best of our knowledge, and according to the recent review of outpatient appointment systems by *Ahmadi-Javid et al.* (2017), this chapter presents the first rigorous and computational analysis of models for SOPSP.

Credits: The results in this chapter are from "Analysis of Models for the Stochastic Outpatient Procedure Scheduling Problem" *Shehadeh et al.* (2019), obtained jointly with Amy E.M Cohn, and Marina A. Epelman.

2.6 Appendix

2.6.1 Comparison of Linear Programming Relaxations of Models (S) of (2.1) and (M) of (2.2)

In this section, we compare the LP relaxations of models (S) of (2.1) and (M) of (2.2) under the assumption that waiting and idling costs are identical across appointments and procedures, i.e., that $\lambda_i^w = \lambda^w$ and $\lambda_i^g = \lambda^g \; \forall i$, and $\lambda_p^w = \lambda^w$ and $\lambda_p^g = \lambda^g \; \forall p$. Since these two models take the same approach to waiting time and cost calculations in case of patient no-shows (see discussion in Section 2.3), we allow for no-shows, which would be represented as procedures with duration 0.

Theorem 2.6.1. *Suppose $\lambda^w > 0$, and $\lambda^g > 0$ and/or $\lambda^o > 0$. The linear programming relaxations of models (S) of (2.1) and (M) of (2.2) are equivalent. In particular, given an optimal solution to the LP relaxation of (S), we can construct a feasible solution to the LP relaxation of (M) with the same objective function value, and vice versa.*

Proof. Suppose $(\hat{x}, \hat{t}, \hat{s}, \hat{g}, \hat{o})$ (with appropriately indexed components) is an optimal solution to the LP relaxation of (S), which is obtained by replacing constraint (2.1k) with $0 \leq \hat{x}_{i,p} \leq 1 \; \forall (i, p)$. Below, we construct a feasible solution $(\bar{x}, \bar{t}, \bar{w}, \bar{g}, \bar{o}, \bar{e})$ to the LP relaxation of (M) with the same objective value. (Recall that components of $\hat{g}$ are indexed differently than those of $\bar{g}$.)

- Let $\bar{x} = \hat{x}$ and $\bar{t} = \hat{t}$. Since $\hat{x}$ satisfies constraints (2.1b) and (2.1c), and $0 \leq \hat{x}_{i,p} \leq 1 \; \forall (i, p)$, $\bar{x}$ satisfies (2.2b) and (2.2c), and $0 \leq \bar{x}_{i,p} \leq 1 \; \forall (i, p)$. Similarly, since $\hat{t}$ satisfies (2.1j) then $\bar{t}$ satisfies (2.2i). Moreover, if symmetry-breaking constraints (3.8) are included in both models, they will be satisfied by both $\hat{x}$ and $\bar{x}$.

- Let $\bar{w}_{i,p}^n = (\hat{s}_i^n - \hat{t}_i) \cdot \hat{x}_{i,p} \; \forall (i,p,n)$. Due to constraints (2.1d), and since $\hat{x}_{i,p} \geq 0$, $\bar{w}_{i,p} \geq 0$ and thus satisfies constraints (2.2h). By construction, $\bar{w}_{i,p} = 0$ whenever $\hat{x}_{i,p} = 0$. Moreover, in an optimal solution of the LP relaxation of (S), $\hat{t}$ and $\hat{s}$ will be chosen to ensure that the values of $\hat{s}_i^n - \hat{t}_i$ will not be excessive for any n as long as $\lambda^w > 0$ (otherwise, one would be able to reduce the waiting component of the cost of the solution). Therefore, constraints (2.2f) will be satisfied for sufficiently large $M_1^i, \; i = 1, \ldots, P$.

- Let $\bar{g}_{i,p}^n = \hat{g}_i^n \cdot \hat{x}_{i,p} \; \forall (i,p,n)$, which clearly satisfies (2.2h). By construction, $\bar{g}_{i,p}^n = 0$ whenever $\hat{x}_{i,p} = 0$. Moreover, in an optimal solution of the LP relaxation of (S), $\hat{t}$ and $\hat{s}$ will be chosen to ensure that the values of $\hat{g}_i^n$ will not be excessive for any n as long as $\lambda^w > 0$, or $\lambda^g > 0$ or $\lambda^o > 0$ (otherwise, one will be able to reduce the waiting or idling/overtime component of the cost of the solution). Therefore, constraints (2.2g) will be satisfies for sufficiently large M_2.

- Let $\bar{o}^n = \hat{o}^n \; \forall n$ (which satisfies (2.2h)), and define $\bar{e}^n$ to satisfy equation (2.2e) $\forall n$.

It remains to verify that the vector $(\bar{x}, \bar{t}, \bar{w}, \bar{g}, \bar{o}, \bar{e})$ defined above satisfies constraints (2.2d), and $\bar{e}^n \geq 0 \; \forall n$.

First, we derive several helpful algebraic expressions. Given the formulae defining $\bar{w}_{i,p}^n$ and $\bar{g}_{i,p}^n$, we have:

$$\sum_{p=1}^{P} \bar{w}_{i,p}^n = \sum_{p=1}^{P} (\hat{s}_i^n - \hat{t}_i) \cdot \hat{x}_{i,p} = (\hat{s}_i^n - \hat{t}_i) \cdot \sum_{p=1}^{P} \hat{x}_{i,p} = \hat{s}_i^n - \hat{t}_i \; \forall (i,n) \tag{2.5}$$

and

$$\sum_{p=1}^{P} \bar{g}_{i,p}^n = \sum_{p=1}^{P} \hat{g}_i^n \cdot \hat{x}_{i,p} = \hat{g}_i^n \cdot \sum_{p=1}^{P} \hat{x}_{i,p} = \hat{g}_i^n \; \forall (i,p), \tag{2.6}$$

where the last equality, in both cases, is due to (2.1c). Using (2.5) and (2.6) and the definition of $\bar{t}$, the left-hand side of (2.2d) can be re-written as

$$\hat{t}_i - \hat{t}_{i+1} - (\hat{s}^n_{i+1} - \hat{t}_{i+1}) + \hat{g}^n_i + (\hat{s}^n_i - \hat{t}_i) = -\hat{s}^n_{i+1} + \hat{g}^n_i + \hat{s}^n_i = -\sum_{p=1}^{P} d^n_p \hat{x}_{i,p} = -\sum_{p=1}^{P} d^n_p \bar{x}_{i,p}, \quad (2.7)$$

where the second equality follows from (2.1f), and the third one — from the definition of $\bar{x}$. This verifies constraints (2.2d).

Finally, using the definition of $\hat{e}^n$ via (2.2e) and expression (2.5), we derive:

$$\bar{e}^n = \bar{o}^n + \mathcal{L} - \sum_{p=1}^{P} d^n_p \bar{x}_{P,p} - \bar{t}_P - \sum_{p=1}^{P} \bar{w}^n_{P,P} = \hat{o}^n + \mathcal{L} - \sum_{p=1}^{P} d^n_p \hat{x}_{P,p} - \hat{s}^n_P \geq 0$$

by (2.1g).

We conclude that $(\bar{x}, \bar{t}, \bar{w}, \bar{g}, \bar{o}, \bar{e})$ defined above is a feasible solution to the LP relaxation of (M), with objective function value

$$\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N}\Big[\sum_{i=1}^{P}\sum_{p=1}^{P}\lambda^w \bar{w}^n_{i,p} + \sum_{i=1}^{P}\sum_{p=1}^{P}\lambda^g \bar{g}^n_{i,p} + \lambda^o \bar{o}^n\Big] \\
=&\frac{1}{N}\sum_{n=1}^{N}\Big[\sum_{i=1}^{P}\sum_{p=1}^{P}\lambda^w (\hat{s}^n_i - \hat{t}_i)\cdot \hat{x}_{i,p} + \sum_{i=1}^{P}\sum_{p=1}^{P}\lambda^g \hat{g}^n_i \cdot \hat{x}_{i,p} + \lambda^o \hat{o}^n\Big] \\
=&\frac{1}{N}\sum_{n=1}^{N}\Big[\sum_{i=1}^{P}\lambda^w (\hat{s}^n_i - \hat{t}_i)\cdot \sum_{p=1}^{P}\hat{x}_{i,p} + \sum_{i=1}^{P}\lambda^g \hat{g}^n_i \cdot \sum_{p=1}^{P}\hat{x}_{i,p} + \lambda^o \hat{o}^n\Big] \\
=&\frac{1}{N}\sum_{n=1}^{N}\Big[\sum_{i=1}^{P}\lambda^w (\hat{s}^n_i - \hat{t}_i) + \sum_{i=1}^{P}\lambda^g \hat{g}^n_i + \lambda^o \hat{o}^n\Big],
\end{aligned}$$

i.e., equal to the optimal value of the LP relaxation of (S).

Conversely, suppose $(\bar{x}, \bar{t}, \bar{w}, \bar{g}, \bar{o}, \bar{e})$ is an optimal solution to the LP relaxation of model (M) of (2.2). We will construct a feasible solution $(\hat{x}, \hat{t}, \hat{s}, \hat{g}, \hat{o})$ to the LP relaxation of (S) with the same objective value.

- Let $\hat{x} = \bar{x}$, $\hat{t} = \bar{t}$, and $\hat{o} = \bar{o}$, which satisfy constraints (2.1b), (2.1c), (2.1i), (2.1j), and

$0 \leq \hat{x}_{i,p} \leq 1 \, \forall (i,p)$. Moreover, if symmetry-breaking constraints (3.8) are included in both models, they will be satisfied by both $\bar{x}$ and $\hat{x}$.

- Let $\hat{s}_i^n = \sum_{p=1}^{P} \bar{w}_{i,p}^n + \bar{t}_i$ and $\hat{g}_i^n = \sum_{p=1}^{P} \bar{g}_{i,p}^n \, \forall (i,n)$. Due to (2.2h), $\hat{s}$ and $\hat{g}$ satisfy (2.1h), and $\hat{s}$ satisfies (2.1d).

With the above definitions, (2.1f) and (2.1e) readily follow from (2.2d) and (2.2h), and (2.1g) follows from (2.2e) and nonnegativity of $\bar{e}$. Therefore, $(\hat{x}, \hat{t}, \hat{s}, \hat{g}, \hat{o})$ is a feasible solution to the LP relaxation of model (S), with objective function value

$$\frac{1}{N}\sum_{n=1}^{N}\Big[\sum_{i=1}^{P}\lambda^w(\hat{s}_i^n - \hat{t}_i) + \sum_{i=1}^{P}\lambda^g\hat{g}_i^n + \lambda^o\hat{o}^n\Big] = \frac{1}{N}\sum_{n=1}^{N}\Big[\sum_{i=1}^{P}\sum_{p=1}^{P}\lambda^w\bar{w}_{i,p}^n + \sum_{i=1}^{P}\sum_{p=1}^{P}\lambda^g\bar{g}_{i,p}^n + \lambda^o\bar{o}^n\Big],$$

i.e., equal to the optimal value of the LP relaxation of (M). This complete the proof. □

Similar analysis techniques can be used to show that the linear programming relaxation of model (S) of (2.1) (and therefore (M) of (2.2)) is at least as tight as the linear programming relaxation of model (B) of (2.3) under the additional assumption that the are no patient no-shows, which needs to be made to account for different approaches to waiting time and cost calculations in these models. Moreover, as illustrated in Table 2.10, linear relaxations of model (S) had larger optimal values, i.e., were tighter, than linear relaxations of model (B) on all test instances in our computational experiments.

CHAPTER 3

Using Stochastic Programming to Solve the Outpatient Appointment Scheduling Problem with Random Service and Arrival Times

3.1 Introduction

Lack of punctuality is a common phenomenon among OPC patients (as well as the customers in many other service industries with an appointment scheduling system), which introduces additional complexity into clinical operations and increases operational costs (*Deceuninck et al.* (2018); *Glowacka et al.* (2017); *Klassen and Yoogalingam* (2014)). Late arrivals often create provider idling (and thus poor utilization of the clinic resources) and, along with random service durations, may increase the waiting time of the subsequent appointments and provider overtime through the propagation of the delay, if there is no buffer in the schedule to absorb it. Very early arrivals, on the other hand, may require the provider to make challenging queuing decisions as whether to serve them ahead of the scheduled time. For example, not serving early arrivals can result in service delay of the next on-time patient (*Deceuninck et al.* (2018); *Glowacka et al.* (2017); *Samorani and Ganguly* (2016)).

Accounting for uncertainity and the possibility of rescheduling (i.e., resequencing or declining) in the scheduling decision process has the potential of mitigating these adverse outcomes. Most of the existing literature on stochastic appointment scheduling (since *Welch and Bailey* (1952)), however, focuses on addressing the variability in service duration only. Within the limited literature accounting for the stochastic arrival time, most studies ignore the sequencing decision and only consider homogeneous patients (i.e., all service durations follow the same distribution) or heterogeneous patients in a sequence found by heuristic approaches (see *Deceuninck et al.* (2018); *Klassen and Yoogalingam* (2014); *Zacharias and Yunes* (2018) and references therein). In addition, most studies adopt the appointment order policy, i.e., the provider will serve the patients in their scheduled appointment order (as opposed to the actual arriving order) and decline a patient if he/she arrives later than the scheduled time by a certain amount of time (termed the *grace period*). Although this policy is easy to implement and effective under many circumstances, to the best of our knowledge, no literature has investigated its optimality (or suboptimality) in a rigorous manner.

In this chapter, we study a stochastic outpatient appointment scheduling problem (SOASP) for OPC scheduling under stochastic arrival times and service durations. We consider an OPC manager who needs to design an appointment schedule and a rescheduling policy for a single provider and a set of patients, where each patient has a known type and associated probability distributions of service duration and arrival time deviation (the difference between the scheduled and the actual arrival times). The objective is to minimize the expected total weighted cost of patient waiting time, provider idle time, and provider overtime.

The problem can be formulated as a *multi-stage* stochastic mixed-integer program (MSMIP) with an initial schedule made in the first stage and rescheduling policy optimized in the subsequent stages. Formulating and solving this MSMIP is challenging, largely because of the enormous size of the scenario tree and the mixed-integer recourse variables (see *Shapiro et al.* (2009); *Birge and Louveaux* (2011) and references therein for a thorough discussion on the difficulty of formulating and solving MSMIP).

In recognition of these challenges, we first consider a *two-stage* model (TSM) that relaxes the non-anticipativity constraints of MSMIP (i.e., the OPC manager has the perfect information of all uncertain parameters when making *re*scheduling decisions) and so yields a lower bound of the optimal value of MSMIP. Second, we derive a family of valid inequalities to strengthen and improve the solvability of this TSM. Third, we obtain an upper bound for MSMIP by solving the TSM model under the appointment order policy. Fourth, we propose a Monte Carlo approach to evaluate the relative gap between the MSMIP upper and lower bounds. Fifth, in a series of numerical experiments, we show that these two bounds are very close in many SOASP parameter settings, demonstrating the near-optimality of the appointment order (AO) policy. We also identify parameter settings that result in a large gap. Accordingly, we close by proposing an alternative swap-based policy that improves the solution in such instances. We also identify parameter settings that result in a large gap in between these two bounds. Accordingly, we propose an alternative policy based on neighbor-swapping. We demonstrate that this alternative policy leads to a much tighter upper bound and significantly shrinks the gap.

The remainder of this chapter is structured as follows. In Section 3.2, we review the relevant literature. In Section 3.3, we formally define SOASP as a MSMIP and present the two TSM approximations. In Section 3.4, we propose a Monte Carlo approach to evaluate the relative gap between the MSMIP upper and lower bounds. In Section 3.5, we report numerical results on (1) the optimality and suboptimality of the appointment order policy and (2) the value of incorporating the stochastic arrival times and service durations in the appointment scheduling problem. Finally, we conclude and summarise this chapter in Section 3.6.

3.2 Literature Review

Most of the stochastic (outpatient) appointment scheduling studies (since the seminal work of *Welch and Bailey* (1952)) assume patient punctuality and focus on homogenous patients. In the following, we review the papers that consider patient heterogeneity and stochastic arrival times.

Cayirli et al. (2006, 2008); *Klassen and Rohleder* (1996); *Rohleder and Klassen* (2000) demonstrated how using patient characteristics, either for appointment sequencing or scheduling, can improve the clinic performance and reduce the operational costs. More recent studies (see *Chen and Robinson* (2014); *Berg et al.* (2014); *Salzarulo et al.* (2016) and references therein) also demonstrated the relevance of recognizing different patient types in appointment sequencing. *Salzarulo et al.* (2016), for example, described how data on patient characteristics and past appointments can be used to predict patient service durations and illustrated how to incorporate this information into effective scheduling decisions.

Patient stochastic arrival (often called patient unpunctuality) is a common phenomenon in OPC and introduces complexity to the design and analysis of appointment scheduling. Several studies have shown that both early and late arrivals disrupt clinic service operations (*Alexopoulos et al.* (2008); *Fetter and Thompson* (1966); *Glowacka et al.* (2017)), increase inefficiencies and delays (*Deceuninck et al.* (2018); *Okotie et al.* (2008)), decrease the service quality for punctual patients *Glowacka et al.* (2017), and increase clinic operational costs (e.g., the provider overtime cost) and implicit costs related to the perceived quality of service (e.g., decreased patient satisfaction) and hence the reputability of the clinic (*Cayirli and Veral* (2003); *Deceuninck et al.* (2018); *Glowacka et al.* (2017); *Kocas* (2015); *Osuna* (1985)). Stochastic arrival times can also lead to adverse operational outcomes and challenging queueing issues (*Deceuninck et al.* (2018); *Cayirli and Veral* (2003); *Jiang et al.* (2019); *Klassen and Yoogalingam* (2014)). For example, not serving an early arrival (or waiting for a chronically late patient) can result in service delay of the next on-time patient (*Deceuninck et al.* (2018); *Glowacka et al.* (2017); *Samorani and Ganguly* (2016)). Therefore, it is worthwhile to design appointment schedules and resequencing policies that mitigate the impacts of both early and late arrivals.

Most empirical studies on stochastic arrival times suggest that patients tend to arrive early instead of being late (*Brahimi and Worthington* (1991); *Cox et al.* (1985); *Fetter and Thompson* (1966); *Klassen and Yoogalingam* (2014); *Klassen and Rohleder* (1996); *Klassen and Yoogalingam*

(2014)). Several studies used the normal, uniform, and exponential (for late patients) distributions to model stochastic arrival deviation (see *Alexopoulos et al.* (2008); *Cheong et al.* (2013); *Cox et al.* (1985); *Tai and Williams* (2012); *White and Pike* (1964) and references therein for a thorough discussion on the distribution of stochastic arrivals). It is worth noting that all of these studies observe that the stochastic arrival time deviation is independent of appointment time. In addition, *Williams et al.* (2014) and *Gneezy et al.* (2011) studied the impacts of strategies that intend to motivate patient punctuality and discussed when and why some strategies might fail. Interventions to reduce tardiness, for example, might potentially lead to increasing patient earliness. Therefore, as pointed out by *Deceuninck et al.* (2018), it is better to incorporate the uncertainty of patient arrival time in the schedule optimization process and develop policies to resequence appointments instead of trying to change the patients' arrival behavior.

Most of the existing literature on stochastic appointment scheduling focuses on the uncertainty pertaining to service durations and/or patient no-shows. Within the few papers accounting for stochastic arrival times, there are three primary approaches. The first approach employs heuristics (often in conjunction with simulation approaches) to develop and adjust the scheduling pattern (see *Cayirli and Yang* (2014); *Cayirli et al.* (2006); *Cox et al.* (1985); *Fetter and Thompson* (1966); *Glowacka et al.* (2017); *Klassen and Yoogalingam* (2014); *White and Pike* (1964), and references therein). Although these pioneering studies provide easy-to-implement scheduling rules, they lack performance guarantee and often ignore the possibility of rescheduling (resequencing and declining). In this chapter, we incorporate rescheduling into our SOASP model and provide computable performance guarantees.

The second approach employs pre-determined resequencing and rescheduling policies. For example, *Deceuninck et al.* (2018) adopted the appointment order policy with 15- and 30-minute grace periods (i.e., the provider will serve patients arriving within the grace period in the scheduled order, and decline or deem as a no-show any late patient coming after the expiration of his/her grace period past the scheduled time). In addition, *Deceuninck et al.* (2018) proposed a local search algorithm based on the Lindley recursion (*Lindley*, 1952) to find an appointment schedule based on a sequence found by a heuristic. Differently, *Zhu et al.* (2017) assumed that the provider will serve a waiting patient regardless of the actual appointment time, and when multiple patients are waiting, the provider will serve them by the order of their appointment times. *Samorani and Ganguly* (2016) studied the problem of whether an available provider should see an early patient right away (preempt) or wait for the next scheduled patient. Based on simulation results, *Samorani and Ganguly* (2016) reported the conditions under which the appointment order policy outperforms the always-preempt policy (e.g., high-service-level clinics with low variability in patient lateness, long service durations, and when patients tend to arrive early rather than late).

First-come-first-serve (FCFS) is another policy which designates that the provider serves the patients according to their actual arriving order, regardless of their initial scheduled order of arrival. This policy often raise the concern that it (partially) conflicts with the goal of appointment scheduling and can encourage the patients to arrive early and beat the appointment system (*Deceuninck et al.* (2018); *Cayirli and Veral* (2003); *Jiang et al.* (2019)). In this chapter, we adopt the appointment order policy to obtain an upper bound on the optimal value of MSMIP. However, we do not assume homogenous patients as in *Samorani and Ganguly* (2016) or a fixed appointment sequence as in *Deceuninck et al.* (2018). Furthermore, we rigorously demonstrate the near-optimality of this policy in a wide range of parameter configurations and propose an alternative and near-optimal policy where this policy is suboptimal.

The third approach employs stochastic mixed-integer programming (SMIP) but, to the best of our knowledge, only involves very few papers (*Ahmadi-Javid et al.* (2017)). *Jiang et al.* (2019) proposed a stochastic linear program that considers stochastic arrival and adopts the appointment order policy. Assuming patient homogeneity, they demonstrated how the optimal appointment intervals change under different stochastic arrivals scenarios. This chapter aims to contribute to this line of research. In contrast to *Jiang et al.* (2019), our SMIP approach considers (1) patient heterogeneity, (2) optimizing both the initial appointment sequencing and scheduling decisions, and (3) the possibility of rescheduling (i.e., resequencing or declining).

3.3 Formulations of SOASP

In this section, we formally define SOASP as a *multi-stage* stochastic mixed-integer program and present two approximations based on *two-stage* stochastic mixed-integer models (TSM). The first approximation relaxes the non-anticipativity constraints of SOASP and yields a lower bound on the optimal value of SOASP. The second approximation applies a (feasible) rescheduling policy and yields an upper bound.

3.3.1 Formal Statement of the Problem

We consider the problem of sequencing a set of patients (appointments) for a single provider and determining the associated scheduled appointment time for each appointment and an adaptive *rescheduling* (i.e., resequencing or declining) policy. Each patient has a known type and a random service duration that follows a known probability distribution associated with the patient type. The actual arrival time of each patient is random relative to their scheduled arrival time, and the arrival time deviation follows a known probability distribution that is independent of the scheduled arrival time.

The provider's response to a patient's arrival depends on three periods (parameters): a *grace period* past the scheduled time, a *rescheduling period* past the grace period, and a *decline period* past the rescheduling period (denoted by intervals G, R, and DE respectively in Figure 3.1). Patients arriving early (i.e., within interval E) or within the grace period (i.e., within interval G) are guaranteed to be served in the scheduled AO. Patients arriving within the rescheduling period (i.e., within interval R) are subject to either resequencing (i.e., the provider will try to fit them within the remaining sequence of appointments if possible) or declining (rescheduled to another day).

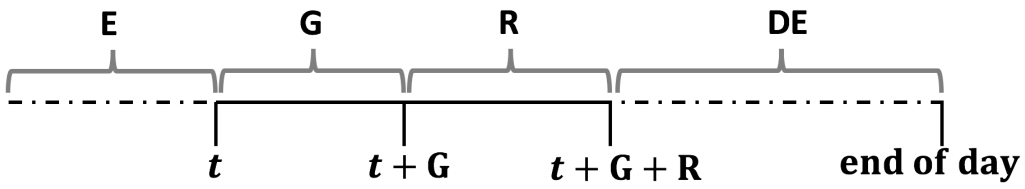


Figure 3.1: An illustration of the provider's response to patient's arrival.
If a patient with scheduled arrival time t arrives any time before $t + G$, then s/he will be served in the scheduled appointment order. If s/he arrives in between $t + G$ and $t + G + R$, then s/he is subject to resequencing or declining. If s/he arrives after $t + G + R$, then s/he will be declined.

Patients arriving within the decline period (i.e., within interval DE) are automatically declined.

The performance metric of scheduling is the weighted sum of three components: (i) *patient waiting time* (the time from the later of the patient *scheduled* arrival time and *actual* arrive time to the actual start time of his/her appointment), (ii) *provider idle time* (the time from the end of one appointment to the start of the next), and (iii) *provider overtime* (time worked beyond the scheduled working hours). Note that we do not incorporate the part of the waiting time due to the early arrival of a patient. That is, if a patient arrives early (i.e., within interval E), his/her waiting time is measured from his/her scheduled time. This is consistent with the prior literature (see, e.g., *Gupta and Wang* (2012); *Cayirli and Veral* (2003)).

Table 3.1 summarises the provider's rescheduling decisions and the corresponding waiting time calculations. We make the following assumptions:

A1. The set of patients (i.e., the total number of appointments and their types) are exogenously determined (a standard assumption in the offline appointment scheduling literature that mimics the OPC practice; see *Ahmadi-Javid et al.* (2017); *Berg et al.* (2014); *Zhu et al.* (2017) and the references therein).

A2. The provider is always available at the start of the day, and immediately after the completion of each appointment (*Deceuninck et al.* (2018); *Berg et al.* (2014)).

Table 3.1: The provider's response to patient's arrival within the intervals defined in Figure 3.1 and waiting time calculations. Notation: AO is appointment order, RES is resequencing, DECL is declining, t is scheduled arrival time, a is actual arrival time, u is arrival time deviation, i.e., $u = a - t$, and s is actual start time.

Arrival	**Interval**	**Arrival time deviation**	**Provider's response**	**Waiting time**
Early or punctual	E	$u \leq 0$	AO	$s - t$
Late but within the grace period	G	$0 < u \leq G$	AO	$s - a$
Late but within the rescheduling period	R	$G < u \leq G + R$	RES or DECL	$s - a$ (RES) or 0 (DECL)
Late but within the decline period	DE	$u > G + R$	DECL	0

A3. The stochastic service duration and arrival time deviation (the difference between the scheduled and the actual arrival times) are independent of the scheduled time (*Berg et al.* (2014); *Deceuninck et al.* (2018); *Denton et al.* (2007); *Mancilla and Storer* (2012)). For the stochastic arrival time deviation, if a patient initially scheduled at 8 AM arrives at 8:20 AM (respectively, 7:40 AM) then the stochastic arrival time deviation equals 20 minutes (respectively, -20 minutes).

A4. The degree of the stochastic arrival time deviation of patient p, u_p, has a bounded support $[\underline{u}_p, \bar{u}_p]$, where $\underline{u}_p$ and $\bar{u}_p$ respectively are the lower and upper bounds on the stochastic arrival time deviation of patient p.

A5. The scheduled times of two consecutive patients are separated by at least one grace period. Note that if we don't make this assumption, it could happen that patient $i + 1$ arrives within the overlap of his/her grace period and that of patient i. This would introduce ambiguity as to whether the provider should take in patient $i+1$ at the arrival time or keep him/her waiting to see if patient i would arrive within his/her grace period.

A6. The grace period and the rescheduling period are nonnegative, i.e., $G \geq 0$ and $R \geq 0$. Furthermore, given assumption A5, $P \times G \leq L$, where P is the number of patients and L the planned length of provider working hours.

SOASP can be formulated as a multi-stage stochastic mixed-integer program (MSMIP). The first stage of the MSMIP pertains to deciding the initial sequence and schedule (i.e., a sequence of scheduled arrival times) of the patients, which are communicated to the patients before the start of the day. In the subsequent stages, the MSMIP implements a policy that reschedules (resequences or declines) the appointments according to the service durations and stochastic arrival times realized up to each new arrival.

Unfortunately, formulating and solving this MSMIP is challenging for two reasons. First, even in a simplified setting in which all random variables are discretized, the scenario tree of the MSMIP can be enormous and grow exponentially in size with the number of patients. For example, if each stage of the MSMIP represents a new patient arrival, then we need multiple random variables to describe the system status at this stage, including which patient the new arrival pertains to, whether this patient arrives in intervals E/G/R/DE, how many existing patients have entered/finished the service, if the provider is currently busy/idle, etc. Second, each stage after the first stage involves a set of binary and continuous decision variables pertaining to the rescheduling (resequencing or declining) of appointments (see *Shapiro et al.* (2009); *Birge and Louveaux* (2011) and references therein for a thorough discussion on the difficulty of formulating and solving MSMIP).

In recognition of the challenges of formulating and solving the MSMIP, we propose two TSMs that respectively lead to a lower bound and an upper bound (and feasible solution) on the optimal value of MSMIP.

3.3.2 TSM under Perfect Information

The first TSM relaxes the non-anticipativity constraints of the MSMIP. That is, the TSM assumes that we possess perfect information (i.e., the realizations of all service durations and stochastic arrival times) after deciding the initial sequence and schedule in the first stage. In this case, the TSM reschedules (i.e., resequences or declines) all appointments based on the perfect information

Table 3.2: Notation of the TSM formulations.

Indices	
p	patient index
i	appointment index in the initial sequence
Parameters	
P	number of patients
c_j^{w}	unit waiting time cost of appointment j
c_j^{g}	unit provider idle time cost after appointment j
c^{o}	unit provider overtime cost
G	grace period
R	rescheduling period
L	planned length of provider working hours
Scenario-dependent parameters	
d_p	realized service duration of patient p
u_p	realized arrival time deviation of patient p
e_p	$\begin{cases} 1, & \text{if patient } p \text{ arrives early (i.e., within interval E) or at the scheduled time,} \\ 0, & \text{otherwise.} \end{cases}$
τ_p	$\begin{cases} 1, & \text{if patient } p \text{ arrives within interval E or G,} \\ 0, & \text{otherwise.} \end{cases}$
δ_p	$\begin{cases} 1, & \text{if patient } p \text{ arrives within the rescheduling period (i.e., interval R),} \\ 0, & \text{otherwise.} \end{cases}$
π_p	$\begin{cases} 1, & \text{if patient } p \text{ arrives within the decline period (i.e, interval DE),} \\ 0, & \text{otherwise.} \end{cases}$
K	number of all patients who can be resequenced or declined , $K \equiv \sum_{p=1}^{P}(\delta_p + \pi_p)$
ξ	random vector containing scenario-dependent parameters, $\xi = (d_1, \ldots, d_P, u_1, \ldots, u_P, e_1, \ldots, e_P, \tau_1, \ldots, \tau_P, \delta_1, \ldots, \delta_P, \pi, \ldots, \pi_P, K)$
Scenario-independent (first stage) variables	
$x_{p,i}$	$\begin{cases} 1, & \text{if patient } p \text{ is initially assigned to appointment } i, \\ 0, & \text{otherwise.} \end{cases}$
t_i	scheduled arrival time of appointment i
Scenario-dependent (second-stage) variables	
$y_{i,j}$	$\begin{cases} 1, & \text{if appointment } i \text{ is resequenced to appointment } j, \\ 0, & \text{otherwise.} \end{cases}$
z_j	$\begin{cases} 1, & \text{if appointment } j \text{ is declined ,} \\ 0, & \text{otherwise.} \end{cases}$
a_j	actual arrival time of appointment j
s_j	actual start time of appointment j
$\widetilde{d}_j$	actual service duration of appointment j
w_j	waiting time of appointment j
g_j	server idle time before the start of appointment j
w_{P+1}	provider overtime

in the second stage. Table 3.2 summarizes the notation that we use in this TSM. Throughout this subsection, we use the term "appointment" and "position" interchangeably to refer to a position in the sequence, and for notational convenience, we suppress the scenario index ξ from the scenario-dependent variables and parameters. We present the TSM under perfect information (TSM-PI) as follows

$$v^{\text{PI}} = \underset{x,t}{\text{minimize}} \quad \mathbb{E}_\xi[Q^{\text{PI}}(x,t,\xi)] \tag{3.1a}$$

$$\text{Subject to: } \sum_{i=1}^{P} x_{p,i} = 1, \qquad \forall p \in [P] \tag{3.1b}$$

$$\sum_{p=1}^{P} x_{p,i} = 1, \qquad \forall i \in [P] \tag{3.1c}$$

$$0 \leq t_i \leq L, \qquad \forall i \in [P] \tag{3.1d}$$

$$t_{i+1} \geq t_i + G, \qquad \forall i \in [P-1] \tag{3.1e}$$

$$x_{p,i} \in \{0,1\}, \qquad \forall (p,i) \in [P] \tag{3.1f}$$

where $[H] := \{h \in \mathbb{N} : 1 \leq h \leq H\}$ for all $H \in \mathbb{N}$ and for each ξ, $Q^{\text{PI}}(x,t,\xi)$ is defined as

$$Q^{\text{PI}}(x,t,\xi) = \underset{a,s,w,g}{\text{minimize}} \quad \sum_{j=1}^{P} \left(c_j^{\text{w}} w_j + c_j^{\text{g}} g_j\right) + c^{\text{o}} w_{P+1} \tag{3.2a}$$

$$\text{Subject to: } z_j \geq \sum_{p=1}^{P} \pi_p x_{p,j}, \qquad \forall j \in [P] \tag{3.2b}$$

$$z_j \leq \sum_{p=1}^{P} (1-\tau_p) x_{p,j}, \qquad \forall j \in [P] \tag{3.2c}$$

$$\sum_{i=1}^{P} y_{i,j} = 1 - z_j, \qquad \forall j \in [P] \tag{3.2d}$$

$$\sum_{j=1}^{P} y_{i,j} = 1 - z_i, \qquad \forall i \in [P] \tag{3.2e}$$

$$\sum_{j=i+1}^{P} y_{i,j} \le \sum_{p=1}^{P} (1-\tau_p) x_{p,i}, \qquad \forall i \in [P-1] \tag{3.2f}$$

$$\sum_{\ell=1}^{j} y_{k\ell} \le \sum_{p=1}^{P} (1-\tau_p) x_{pi} - y_{ij} + 1, \quad \forall 1 \le j \le i \le P, \forall k \in [i+1, P]_{\mathbb{Z}} \tag{3.2g}$$

$$\widetilde{d}_j = \sum_{p=1}^{P} \sum_{i=1}^{P} d_p x_{p,i} y_{i,j}, \qquad \forall j \in [P] \tag{3.2h}$$

$$a_j = \sum_{i=1}^{P} \left(t_i + \sum_{p=1}^{P} u_p x_{p,i} \right) y_{i,j}, \qquad \forall j \in [P] \tag{3.2i}$$

$$s_j \ge a_j, \qquad \forall j \in [P] \tag{3.2j}$$

$$s_j \ge s_{j-1} + \widetilde{d}_{j-1}, \qquad \forall j \in [2, P]_{\mathbb{Z}} \tag{3.2k}$$

$$w_j \ge s_j - a_j - M_j \left(z_j + \sum_{p=1}^{P} \sum_{i=1}^{P} e_p x_{p,i} y_{i,j} \right), \qquad \forall j \in [P] \tag{3.2l}$$

$$w_j \ge s_j - \sum_{i=1}^{P} t_i y_{i,j} - M_j \left(2 - (1 - z_j) - \sum_{p=1}^{P} \sum_{i=1}^{P} e_p x_{p,i} y_{i,j} \right), \quad \forall j \in [P] \tag{3.2m}$$

$$g_1 = s_1 \tag{3.2n}$$

$$g_j = s_j - \left(s_{j-1} + \widetilde{d}_{j-1} \right), \qquad \forall j \in [2, P]_{\mathbb{Z}} \tag{3.2o}$$

$$w_{P+1} \ge s_P + \widetilde{d}_P - L, \tag{3.2p}$$

$$y_{i,j} \in \{0, 1\}, \qquad \forall (i, j) \in [P] \tag{3.2q}$$

$$(g_j, s_j, w_j, w_{P+1}) \ge 0, \qquad \forall j \in [P] \tag{3.2r}$$

where $[2, P]_{\mathbb{Z}} := \{j \in \mathbb{N} : 2 \le j \le P\}$ for all $P \in \mathbb{N}$. Objective (3.1a) minimizes the expected total costs of patient waiting, provider idle time, and provider overtime. Constraints (3.1b) and (3.1c) ensure that initially each patient is assigned to one appointment and each appointment is assigned one patient, respectively. Constraints (3.1d) ensure that all scheduled appointments are within the provider's regular service hours $[0, L]$. Constraints (3.1e) ensure that the scheduled appointment times of consecutive appointments are at least separated by the grace period (see assumptions A5–A6).

Recall that in TSM-PI we assume that we possess perfect information about the stochastic arrival time deviations and service durations. For modeling convenience, we choose to treat any patient with appointment declined (whether he/she arrives in interval DE, and thus is automatically declined, or in interval R and the model declines him/her) as a "ghost" patient arriving at time 0 (this ensures that the provider does not wait for his/her grace period to expire before moving on to the next patient), and with zero service duration. We also force the waiting time of the "ghost" patient to be zero, as explained below.

Constraints (3.2b)–(3.2g) determine the new sequence. Specifically, constraints (3.2b) specify that if a patient p initially assigned to position j (i.e., $x_{p,j} = 1$) arrives within the decline period (within interval DE and so $\pi_p = 1$) then this patient (and the corresponding appointment j) must be declined. Constraints (3.2c) specify that if a patient p initially assigned to position j arrives any time before the end of the grace period (i.e., within interval E or G and so $\tau_p = 1$), then this patient (and the corresponding appointment) must not be declined. Constraints (3.2d) specify that if a patient initially assigned to appointment j is declined, then no appointment can be resequenced to appointment j, i.e., $y_{i,j} = 0 \ \forall i$ (recall that we still "treat" patient j, as a "ghost" with zero service duration). Constraints (3.2e) specify that if the initial appointment i is declined, then this appointment cannot be resequenced to any other appointment.

Constraints (3.2f) specify that if a patient p initially assigned to appointment i arrives early or within the grace period (i.e., within interval E or G and so $\tau_p = 1$ by definition), then he/she cannot be resequenced to an appointment later than i. Constraints (3.2g) specify that if a patient initially assigned to appointment i arrives early or within the grace period, then any patients initially scheduled after him/her cannot be resequenced to an appointment before him/her. Note that, under this global resequencing rule, if the patient initially assigned to appointment i arrives within rescheduling period (i.e., interval R) and doesn't get declined, then he/she can still be resequenced to a position before i. For example, suppose that patients 1 and 2 arrive within interval E or G and patients 3 and 4 arrive within interval R. We are allowed to resequence them as 1–2–4–3 in this case. However, neither 3 nor 4 can be resequenced before 1 or 2. This is reflected in constraints (3.2g). Note that constraints (3.2d)–(3.2g) allow for the initial and the new positions to be the same. In this case, $j = i$ and so $y_{i,i} = 1$.

Constraints (3.2h) determine the actual service duration of appointment j based on the new sequence: if appointment j is declined (i.e., if $z_j = 1$ and so $y_{i,j} = 0$ for all i by constraints (3.2d)) then $\widetilde{d}_j = 0$; and if appointment j is not declined (i.e., if $z_j = 0$ and so $y_{i,j} = 1$ for some i) then $\widetilde{d}_j = d_p$ if patient p is initially assigned to appointment i and then resequenced to appointment j (i.e., $x_{p,i} = 1$ and $y_{i,j} = 1$). Note that constraints (3.2h) and some of the remaining constraints of this TSM are nonlinear because of the bilinear terms $x_{p,i}y^n_{i,j}$ (see constraints (3.2h)–(3.2i) and (3.2l)–(3.2m)) and $t_i y^n_{i,j}$ (see constraints (3.2i)) In section 3.4.2, we propose strategies to linearize these constraints.

Constraints (3.2i) determine the actual arrival time. Specifically, if patient p is initially assigned to position i (i.e., $x_{p,i} = 1$) and resequenced to position j (i.e., $y_{i,j} = 1$ and $z_j = 0$) then the actual arrival time of appointment j is $a_j = t_i + u_p$. In contrast, if appointment j is declined (i.e., $z_j = 1$ and so $y_{i,j} = 0$ for all $i = 1, \dots, P$ by (3.2d)) then $a_j = 0$ (recall that this is a “ghost” patient arriving at time zero by design). Constraints (3.2j)–(3.2k) specify that the actual start time of appointment j, s_j, is no smaller than the actual arrival time, a_j, and no smaller than the completion time of the preceding appointment, i.e., $s_{j-1} + \widetilde{d}_{j-1}$. Note that if appointment j is declined then the corresponding “ghost” patient will start and complete at the completion time of the preceding appointment $s_{j-1} + \widetilde{d}_{j-1}$ because $a_j = 0$ and $\widetilde{d}_j = 0$ for all declined appointments by (3.2i) and (3.2h).

Constraints (3.2l)–(3.2m) compute the waiting time based on the following four arrival time scenarios (see the first and last columns of Table 3.1) with M_j representing a sufficiently large constant. First, if patient p is initially assigned to appointment i, arrives early or at the scheduled time (i.e., within interval E), and is resequenced to appointment j (i.e., $x_{p,i} = y_{i,j} = 1$), then $\tau_p = 1$

by definition and so $z_j = 0$ by (3.2c), $e_p = 1$ by definition, and $\sum_{p=1}^{P}\sum_{i=1}^{P} e_p x_{p,i} y_{i,j} = 1$. It follows that constraint (3.2l) is relaxed and $w_j = \max\{s_j - t_i, 0\}$ by (3.2m) and (3.2r), i.e., the waiting time of this patient is the difference between his/her actual start time and the scheduled arrival time. Recall from Section 3.3.1 that, consistent with the prior literature, we do not incorporate the part of the waiting time due to the early arrival of a patient (see, e.g., *Gupta and Wang* (2012); *Cayirli and Veral* (2003)).

Second, if patient p is initially assigned to appointment i arrives late but within the grace period (i.e., interval G), and is resequenced to appointment j, then $\tau_p = 1$ by definition and so $z_j = 0$ by (3.2c), $e_p = 0$ by definition, $\sum_{p=1}^{P}\sum_{i=1}^{P} e_p x_{p,i} y_{i,j} = 0$, and $a_j = t_i + u_p$ by (3.2i). It follows that constraint (3.2m) is relaxed and $w_j = s_j - a_j$ by (3.2l), i.e., the waiting time of this late patient is computed from his/her actual arrival time. Third, if patient p is initially assigned to appointment i arrives late but within the rescheduling period (i.e., interval R) but does not get declined, and resequenced to appointment j, then $z_j = 0$, $a_j = t_i + u_p$ by (3.2i), $e_p = 0$ by definition, constraint (3.2m) is relaxed, and $w_j = s_j - a_j$ by (3.2l). Fourth, if patient p initially assigned to appointment j arrives within the decline period (i.e., interval DE), then $\pi_p = 1$ by definition and so $z_j = 1$ by (3.2b), $y_{i,j} = 0\ \forall i$ by (3.2d), and $\sum_{p=1}^{P}\sum_{i=1}^{P} e_p x_{p,i} y_{i,j} = 0$. It follows that both (3.2l)–(3.2m) are relaxed and $w_j = 0$ by (3.2r), i.e., declined appointments yield zero waiting time.

Constraint (3.2n) computes the idle time before the first patient. Constraints (3.2o) computes the idle time between two consecutive appointments. Constraints (3.2p) and (3.2r) compute the provider overtime (if any) beyond the planned length of working hours. Finally, constraints (3.2q)–(3.2r) specify binary and nonnegative restrictions on the decision variables.

Proposition 3.3.1. *For a fixed grace period G and cost vectors c^{w}, c^{g}, and c^{o}, we have $v^{PI} \leq v$, where v is the optimal value of the MSMIP.*

Proof. Formulation (3.2) relaxes the nonanticaptivity constraints in the MSMIP. It follows that $v \geq v^{\text{PI}}$. □

3.3.3 TSM under the Appointment Order Policy

The second TSM implements a fixed (feasible) rescheduling policy termed *appointment order policy* (AO), which designates that (i) any patient arriving after his/her grace period (i.e., within interval R or DE) gets declined and (ii) the remaining patients are served in the order of their initial sequence regardless of their actual arrival times. Using the same notation in Table 3.2, we present the TSM under AO (TSM-AO) as follows.

$$v^{\text{AO}} = \underset{x,t}{\text{minimize}} \ \mathbb{E}_{\xi}[Q^{\text{AO}}(x,t,\xi)] \tag{3.3a}$$

$$\text{Subject to: (3.1b)} - \text{(3.1f)} \tag{3.3b}$$

where for each ξ, $Q^{\text{AO}}(x,t,\xi)$ is defined as

$$Q^{\text{AO}}(x,t,\xi) = \underset{a,s,w,g}{\text{minimize}} \quad \sum_{j=1}^{P} \left(c_j^{\text{w}} w_j + c_j^{\text{g}} g_j\right) + c^{\text{o}} w_{P+1} \tag{3.3c}$$

$$\text{Subject to: } a_j = t_j + \sum_{p=1}^{P} \big(u_p \tau_p + G(1-\tau_p)\big) x_{p,j}, \qquad \forall j \in [P] \tag{3.3d}$$

$$s_j \geq a_j, \qquad \forall j \in [P] \tag{3.3e}$$

$$s_j \geq s_{j-1} + \sum_{p=1}^{P} d_p \tau_p x_{p,j-1}, \qquad \forall j \in [2,P]_{\mathbb{Z}} \tag{3.3f}$$

$$w_j \geq s_j - a_j - M_j \left(1 + \sum_{p=1}^{P} (e_p - \tau_p) x_{p,j}\right), \qquad \forall j \in [P] \tag{3.3g}$$

$$w_j \geq s_j - t_j - M_j \left(2 - \sum_{p=1}^{P} (e_p + \tau_p) x_{p,j}\right), \qquad \forall j \in [P] \tag{3.3h}$$

$$g_1 = s_1, \tag{3.3i}$$

$$g_j = s_j - \left(s_{j-1} + \sum_{p=1}^{P} d_p \tau_p x_{p,j-1}\right), \qquad \forall j \in [2,P]_{\mathbb{Z}} \tag{3.3j}$$

$$w_{P+1} \geq s_P + \sum_{p=1}^{P} d_p \tau_p x_{p,P} - L, \tag{3.3k}$$

$$(g_j, s_j, w_j, w_{P+1}) \geq 0, \qquad \forall j \in [P] \tag{3.3l}$$

Note that under the AO policy, the provider must wait for a patient till the end of his/her grace period before becoming eligible to treat a subsequent patient (Deceuninck et al., 2018). For modeling convenience, we choose to treat any patient arriving after the end of the grace period (i.e., within interval R or DE) as a "ghost" patient arriving at $t + G$ with zero service duration.

Objective (3.3a) minimizes the expected total costs of patient waiting, provider idle time, and provider overtime. Constraints (3.3d) determine the actual arrival time. Specifically, if patient p assigned to appointment j (i.e., $x_{p,j} = 1$) arrives early or within the grace period (i.e., within interval E or G) and so $\tau_p = 1$ by definition, then $a_j = t_j + u_p$. In contrast, if this patient arrives after the end of the grace period (i.e, within interval R or DE) and so $\tau_p = 0$ by definition, then this patient is declined and $a_j = t_j + G$, i.e., this patients is treated as a "ghost" patient arriving at $t_j + G$ with zero service time, i.e., $\sum_{p=1}^{P} d_p \tau_p x_{p,j} = 0$.

Constraints (3.3e)–(3.3f) ensure that the actual start time is at least the arrival time of the patient and at least the completion time of the previous patient. Note that if appointment j is declined then the corresponding "ghost" patient p will start and complete at $s_j + \sum_{p=1}^{P} d_p \tau_p x_{p,j} = \max\{a_j = t_j + G, s_{j-1} + \sum_{p=1}^{P} d_p \tau_p x_{p,j-1}\} + 0$ because $\tau_p = 0$ and so $\sum_{p=1}^{P} d_p \tau_p x_{p,j} = 0$. Collectively, constraints (3.3d)–(3.3f) enforce the AO policy mathematically. For example, if patient 4 arrives earlier than $t_3 + G$ (i.e., the expiration time of patient 3's grace period), the *provider whenever becomes available will remain idle until one of the following two cases takes place*: (1) patient 3 arrives before $t_3 + G$ and the provider starts serving patient 3, or (2) $t_3 + G$ expires, and the provider declines patient 3 and starts serving patient 4.

Constraints (3.3g)–(3.3h) compute the waiting time based on the following three arrival time scenarios with M_j representing a sufficiently large constant. First, if patient p is assigned to appointment j (i.e., $x_{p,j} = 1$) and arrives early or at the scheduled time (i.e., within interval E), then $e_p = 1$ and $\tau_p = 1$ by definition, and thus $\sum_{p=1}^{P}(e_p - \tau_p)x_{p,j} = 0$, and $\sum_{p=1}^{P}(e_p + \tau_p)x_{p,j} = 2$. It follows that constraint (3.3g) is relaxed and $w_j = \max\{s_j - t_j, 0\}$ by constraints (3.3h) and (3.3l). Second, if patient p is assigned to appointment j (i.e., $x_{p,j} = 1$) and arrives within the grace period (i.e., within interval G), then $e_p = 0$ and $\tau_p = 1$ by definition, and so $\sum_{p=1}^{P}(e_p - \tau_p)x_{p,j} = -1$ and $\sum_{p=1}^{P}(e_p + \tau_p)x_{p,j} = 1$. It follows that constraint (3.3h) is relaxed and $w_j = s_j - a_j$ by constraint (3.3g). Third, if patient p assigned to appointment j arrives after the end of the grace period (i.e, within interval R or DE), then $e_p = \tau_p = 0$ by definition and so $\sum_{p=1}^{P}(e_p - \tau_p) = \sum_{p=1}^{P}(e_p + \tau_p)x_{p,j}$=0. It follows that both (3.3g) and (3.3h) are relaxed and $w_j = 0$ by (3.3l), i.e., declined appointments yield zero waiting time.

Constraint (3.3i) computes the idle time before the first patient. Constraints (3.3j) compute the idle time between two consecutive appointments. Constraints (3.3k) and (3.3l) compute the provider overtime beyond the planned length of working hours. Finally, constraints (3.3l) specify nonnegative restrictions on the decision variables.

Proposition 3.3.2. *For a fixed grace period G and cost vectors c^w, c^g, and c^o, we have $v \leq v^{AO}$. Furthermore, $v^{PI} \leq v \leq v^{AO}$.*

Proof. The appointment order policy is a feasible rescheduling policy to the MSMIP. It follows that $v \leq v^{\text{AO}}$, and so $v^{\text{PI}} \leq v \leq v^{\text{AO}}$ by Proposition 3.3.1. □

3.4 Solution Approach

There are two well-known difficulties in obtaining an (exact) optimal solution to the TSM-PI in (3.1) and TSM-AO in (3.3). First, evaluating the values of $\mathbb{E}_\xi[Q^{\text{PI}}(x,t,\xi)]$ and $\mathbb{E}_\xi[Q^{\text{AO}}(x,t,\xi)]$ involves taking multi-dimensional integrals. Second, formulation (3.2) of $Q^{\text{PI}}(x,t,\xi)$ involves mixed-integer recourse variables, which makes $Q^{\text{PI}}(x,t,\xi)$ non-convex and even discontinuous.

In view of these two difficulties, we resort to approximation solution approaches.

In Section 3.4.1, we present a Monte Carlo approach to obtain near-optimal solutions to (3.1) and (3.3). In Section 3.4.2, we propose several strategies to linearize and strengthen the formulation (3.1), which helps improve the solution efficacy of the TSM-PI.

3.4.1 Monte Carlo Optimization

In the Monte Carlo approach, we replace the distribution of ξ with a (discrete) empirical distribution based on N independent and identically distributed (i.i.d.) samples of the service durations and arrival time deviations, and then we solve the SAA formulations of (3.1) and (3.3), which are denoted formulations (3.4) and (3.5), respectively, and presented below. Note that, in the SAA formulations (3.4)–(3.5), we associate all scenario-dependent parameters, variables, and constraints with a scenario index n for all $n = 1, \dots, N$. For example, parameters d_p are replaced by d_p^n to represent the service durations realized in scenario n, and variables $y_{i,j}$ are replaced by $y_{i,j}^n$ to represent the resequencing decisions in scenario n. In addition, constraints (3.2b)–(3.2r) and (3.3d)–(3.3l) are incorporated in each scenario.

$$\text{(SAA-PI)}\ \ v_N^{\text{PI}} = \underset{x,t,a,s,w,g}{\text{minimize}}\ \hat{f}_N^{\text{PI}} := \frac{1}{N} \sum_{n=1}^{N} \sum_{j=1}^{P} \left(c_j^{\text{w}} w_j^n + c_j^{\text{g}} g_j^n \right) + c^{\text{o}} w_{P+1}^n \tag{3.4a}$$

$$\text{Subject to:}\ \ (3.1\text{b}) - (3.1\text{f}) \tag{3.4b}$$

$$(3.2\text{b}) - (3.2\text{r}),\ \ \forall n \in [N] \tag{3.4c}$$

$$\text{(SAA-AO)}\ \ v_N^{\text{AO}} = \underset{x,t,a,s,w,g}{\text{minimize}}\ \hat{f}_N^{\text{AO}} := \frac{1}{N} \sum_{n=1}^{N} \sum_{j=1}^{P} \left(c_j^{\text{w}} w_j^n + c_j^{\text{g}} g_j^n \right) + c^{\text{o}} w_{P+1}^n \tag{3.5a}$$

$$\text{Subject to: } (3.1\text{b}) - (3.1\text{f}) \tag{3.5b}$$

$$(3.3\text{d}) - (3.3\text{l}), \quad \forall n \in [N] \tag{3.5c}$$

Note that the sample averages $\hat{f}_N^{\text{PI}}$ and $\hat{f}_N^{\text{AO}}$ are unbiased estimators of the expected values $f^{\text{PI}} := \mathbb{E}_\xi[Q^{\text{PI}}(x,t,\xi)]$ and $f^{\text{AO}} := \mathbb{E}_\xi[Q^{\text{AO}}(x,t,\xi)]$ in (3.1) and (3.3), respectively (see *Shapiro* (2003); *Mak et al.* (1999) and references therein). By the Law of Large Numbers and *Shapiro* (2003), we have $\hat{f}_N^{\text{PI}} \to f^{\text{PI}}$ and $\hat{f}_N^{\text{AO}} \to f^{\text{AO}}$ with probability one (w.p.1) as $N \to \infty$ (*Linderoth et al.* (2006); *Homem-de Mello and Bayraksan* (2014); *Kleywegt et al.* (2002)). It follows that $v_N^{\text{PI}} \to v^{\text{PI}}$ and $v_N^{\text{AO}} \to v^{\text{AO}}$ w.p.1 as $N \to \infty$, i.e., the optimal values of the SAA formulations (3.4) and (3.5) converge to those of TSM-PI and TSM-AO, respectively, as the sample size N grows to infinity. However, for a fixed sample of N scenarios, formulation (3.4) and (3.5) reduce to a mixed-integer nonlinear program (MINLP) and a mixed-integer linear program (MILP), respectively. Hence, one would expect the computational effort and solution time of solving the SAA formulations to increase as the sample size increases.

Algorithm 3.1 summarizes the Monte-Carlo Optimization (MCO) algorithm that determines an appropriate sample size N and obtains near-optimal solutions to the TSM-PI model based on SAA formulations, and the algorithm for solving the TSM-AO model is similar. This algorithm is based on the vanilla SAA method in *Ahmed et al.* (2002), *Homem-de Mello and Bayraksan* (2014), *Kleywegt et al.* (2002), and *Molina-Pariente et al.* (2016) with some adaptations to our TSM models.

Starting with an initial candidate value of N, the MCO algorithm proceeds as follows. First, for $m = 1, \dots, M$, we repeat the following steps. In step 1.1, we generate a sample of N i.i.d. scenarios of service durations and arrival time deviations. In step 1.2, we solve the SAA formulation of the TSM-PI with the scenarios generated in step 1.1 and record the corresponding optimal objective value v_N^m and optimal schedule $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{t}})_N^m$. In step 1.3, we evaluate the objective function value $v_{N'}^m$ via Monte Carlo simulation of the schedule $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{t}})_N^m$ with a sample of N' i.i.d. scenarios of service durations and arrival time deviations.

Algorithm 3.1: Monte Carlo Optimization (MCO) Method

```
Input: $N_o$ is an initial sample size, $M$ is number of replicates, $N'$ is number of scenarios in the Monte Carlo
       Simulation step, and $\epsilon$ is a termination tolerance.
Output: $N$ is sample size, $\overline{v}_N$ and $\overline{v}_{N'}$ are respectively statistical lower and upper bounds on the optimal value
        of the TSM, and $AOI_N$ is approximate optimality index.
Initialization: $N := N_o$

Step 1. MCO Procedure
for $m = 1, ..., M$, do
    Step 1.1 Scenario Generation
        - Generate $N$ independent and identical distributed (i.i.d.) scenarios of service durations
          $\mathbf{d}_p = (d_p^1, ...., d_p^N)^T$ and arrival time deviations $\mathbf{u}_p = (u_p^1, \ldots, u_p^N)^T$ for all $p = 1, ..., P$.
    Step 1.2 Solving the SAA formulation
        - Solve the SAA formulation in (3.4) with the scenarios generated in step 1.1 and record the
          corresponding optimal objective value $v_N^m$ and optimal schedule $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{t}})_N^m$.
    Step 1.3 Cost Evaluation using Monte Carlo Simulation
        - Generate $N'$ i.i.d. scenarios of service durations $\mathbf{d}'_p = (d_p^1, ...., d_p^{N'})^T$ and arrival time
          deviations $\boldsymbol{u}'_p = (u_p^1, \ldots, u_p^{N'})^T$ for all $p = 1, ..., P$.
        - Use the schedule $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{t}})_N^m$ and parameters $(\mathbf{d}'_p, \boldsymbol{u}'_p)$ to compute $w_j^{n'}$, $g_j^{n'}$, and $w_{P+1}^{n'}$ for all
          $n' = 1, \ldots, N'$, and evaluate the objective function $\hat{v}_{N'}^m$ as follows:

          $$\hat{v}_{N'}^m = \frac{1}{N'} \sum_{n'=1}^{N'} \sum_{j=1}^{P} \left( c_j^{\mathrm{w}} w_j^{n'} + c_j^{\mathrm{g}} g_j^{n'} \right) + c^{\mathrm{o}} w_{P+1}^{n'}$$
end
Step 2. Compute the average of $\hat{v}_N^m$ and $\hat{v}_{N'}^m$ among the M replications

    $$\overline{v}_N = \frac{1}{M} \sum_{m=1}^{M} v_N^m \qquad \overline{v}_{N'} = \frac{1}{M} \sum_{m=1}^{M} \hat{v}_{N'}^m$$
Step 3. Compute the Approximate Optimality Indix

    $$AOI_N = \frac{\overline{v}_{N'} - \overline{v}_N}{\overline{v}_{N'}}$$
Step 4. If $AOI_N$ satisfies a predetermined termination tolerance (i.e., $|AOI_N| < \epsilon$), terminate and output $N$,
$\overline{v}_N$, $\overline{v}_{N'}$, and $AOI_N$. Otherwise, update $N \leftarrow 2N$, and go to step1.
```

In step 2, we compute the average of v_N^m and $v_{N'}^m$ among the M replications as $\overline{v}_N = M^{-1} \sum_{m=1}^{M} v_N^m$ and $\overline{v}_{N'} = M^{-1} \sum_{m=1}^{M} \hat{v}_{N'}^m$, respectively. The statistical results in Mak et al. (1999) and Linderoth et al. (2006) infer that $\overline{v}_N$ and $\overline{v}_{N'}$ are respectively statistical lower and upper bounds of the optimal value of the TSM model. In step 3, we compute the approximate optimality index $|AOI_N = (\overline{v}_{N'} - \overline{v}_N)/\overline{v}_{N'}|$ as a point estimate of the relative optimality gap between these two statistical bounds, and so AOI_N serves as an approximate estimate of the optimality gap. Finally, if AOI_N satisfies a predetermined termination tolerance (i.e., $|AOI_N| < \epsilon$), the algorithm terminate and output N, $\overline{v}_N$, $\overline{v}_{N'}$, and AOI_N. Otherwise, we increase the sample size (i.e., $N \leftarrow 2N$), and go to step 1.

Finally, we can "warm start" Algorithm 3.1 from solved instances. That is, we can store the output sample size N in a solved instance (i.e., N leads to a near-optimal solution to the TSM via its SAA formulation) and set the initial sample size N_o=N when solving similar instances. This can significantly speed up Algorithm 3.1.

We evaluate the gap between the two TSM approximations of SOASP, i.e., Gap $:= \frac{v^{\text{AO}} - v^{\text{PI}}}{v^{\text{PI}}}$, where v^{PI} and v^{AO} represent the optimal values of TSM-PI and TSM-AO, respectively. As $\bar{v}_N^{\text{PI}}$ is a statistical lower bound of v^{PI} and $\overline{v}_{N'}^{\text{AO}}$ is a statistical upper bound of v^{AO}, we conservatively approximate Gap by using the approximate (statistical) relative gap (ARG), where ARG $:= \frac{\overline{v}_{N'}^{\text{AO}} - \overline{v}_N^{\text{PI}}}{\overline{v}_N^{\text{PI}}}$.

3.4.2 Strengthening the TSM-PI

As compared to the TSM-AO, the SAA formulation (3.4) of the TSM-PI is significantly more difficult to solve for three main reasons. First, the formulation (3.4) is a mixed-integer nonlinear programming (MINLP) because of the bilinear terms $x_{p,i}y_{i,j}^n$ (see constraints (3.2h)–(3.2i) and (3.2l)–(3.2m)) and $t_i y_{i,j}^n$ (see constraints (3.2i) and (3.2m)). Second, as we consider rescheduling (resequencing and declining) in the TSM-PI, the formulation (3.4) involves integer recourse variables $y_{i,j}^n$ and z_j^n. Third, the search space of the rescheduling variables increases with the sample size (i.e., the number of binary variables $y_{i,j}^n$ and z_j^n grows linearly with the sample size N). In the following subsections, we strengthen the formulation (3.4) by linearization and valid inequalities.

3.4.2.1 SMILP Reformulation

Let $\mathcal{N}^{\text{EG}}$ be the set of all scenarios in which all patients arrive early or within the grace period (i.e., within interval E or G and so $\tau_p^n = 1$ and $\delta_p^n = 0,$ for all $p = 1, \dots, P$ and $n \in \mathcal{N}^{\text{EG}}$). In this set of scenarios, the provider must follow the initial appointment order and not resequence any patient. Accordingly, we have $y_{j,j}^n = 1$ and $z_j^n = 0$ for all $j = 1, \dots, P$ and $n \in \mathcal{N}^{\text{EG}}$. Therefore, the nonlinear constraints (3.2h)–(3.2p) reduce to linear constraints (3.3d)–(3.3k) for all $n \in \mathcal{N}^{\text{EG}}$.

To linearize formulation (3.4) for all other scenarios $n \notin \mathcal{N}^{\text{EG}}$, we define $\alpha_{i,j}^n = t_i y_{i,j}^n,\ \forall (i,j) \in [P]$ and $\beta_{p,i,j}^n = x_{p,i} y_{i,j}^n, \forall (p,i,j) \in [P]$ (where $[H] := \{h \in \mathbb{N} : 1 \leq h \leq H\}$ for all $H \in \mathbb{N}$). We also incorporate the following McCormick inequalities (3.6a)–(3.6b) and (3.6c)–(3.6d) for variables $\alpha_{i,j}^n$ and $\beta_{p,i,j}^n$, respectively.

$$\alpha_{i,j}^n - t_i^{\text{min}} y_{i,j}^n \geq 0, \qquad \alpha_{i,j}^n - t_i + t_i^{\text{max}}(1 - y_{i,j}^n) \geq 0 \tag{3.6a}$$

$$\alpha_{i,j}^n - t_i^{\text{max}} y_{i,j}^n \leq 0, \qquad \alpha_{i,j}^n - t_i + t_i^{\text{min}}(1 - y_{i,j}^n) \leq 0 \tag{3.6b}$$

$$\beta_{p,i,j}^n \geq 0, \qquad \beta_{p,i,j}^n - y_{i,j}^n - x_{p,i} + 1 \geq 0 \tag{3.6c}$$

$$\beta_{p,i,j}^n - x_{p,i} \leq 0, \qquad \beta_{p,i,j}^n - y_{i,j}^n \leq 0 \tag{3.6d}$$

where t_i^{min} and t_i^{max} are the minimum and maximum possible values of the scheduled time of appointment i, respectively. We derive a tight estimation of t_i^{min} and t_i^{max} in proposition 3.4.1, and we relegate the proof to Appendix 3.7.1. We also relegate the resulting stochastic mixed-integer linear program (SMILP) to Appendix 3.7.2.

Proposition 3.4.1. *$t_i^{min} \leq t_i \leq t_i^{max}$, where $t_i^{min} = (i-1)G$ and $t_i^{max} = L - (P-i)G$ for all $i \in [P]$.*

3.4.2.2 Valid Inequalities

We derive two families of valid inequalities to strengthen formulation (3.4). We summarize these inequalities in the following proposition.

Proposition 3.4.2.

$$\sum_{j=\max\{i-K^n,1\}}^{i} y_{i,j}^n \geq \sum_{p=1}^{P} \tau_p^n x_{p,i}, \qquad \forall n \in [N],\ \forall i \in [P], \tag{3.7a}$$

$$y_{i,i}^n \geq 1 - \sum_{k=1}^{i} \sum_{p=1}^{P} (1-\tau_p^n) x_{pk}, \qquad \forall n \in [N],\ \forall i \in [P]. \tag{3.7b}$$

where $K^n \equiv \sum_{p=1}^{P} (\delta_p^n + \pi_p^n)$.

Proof. Inequalities (3.7a) specify that if a patient p initially assigned to appointment i arrives early or within the grace period in scenario n (i.e., within interval E or G and so $\tau_p^n = 1$ by definition), then he/she can only be resequenced to an appointment j among $\max\{i - K^n, 1\}, \ldots, i$. This is because in each scenario n there are at most K^n many appointments before i that can either be declined or be resequenced to be after i. Inequalities (3.7b) specify that if all patients in the first i appointments arrive early or within the grace period, then appointment i must stay as i. This is because (1) the patient in appointment i cannot swap with any patients before him/her as none of them arrive after the grace period, and (2) he/she cannot swap with any patients after him/her as he/she arrives before the start of the rescheduling period.

□

3.4.2.3 Tight Estimation of Big-M Coefficients

We derive a tight estimation of the Big-M coefficients involved in constraints (3.2l)–(3.2m) and (3.3g)–(3.3h) in the following proposition, whose proof is relegated to Appendix 3.7.3. The tight estimation strengthens the SAA formulations of both TSM-PI and TSM-AO.

Proposition 3.4.3. $M_j^n = L + \max_{p=1,\ldots,P}\{|u_p^n|\} + (j-1)\max_{p=1,\ldots,P}\{d_p^n\}$ *for* $j = 1, \ldots, P$ *and* $n \in [N]$ *is a valid value for the Big-M constant in constraints* (3.2l)–(3.2m) *and* (3.3g)–(3.3h).

3.4.2.4 Symmetry Breaking Inequalities

We apply symmetry breaking inequalities (see *Denton et al.*, 2010; *Berg et al.*, 2014; *Shehadeh et al.*, 2019; *Ostrowski et al.*, 2011) to further strengthen the TSM formulations. Specifically, we aggregate the patients into classes, each consisting of patients having a common patient type and a common distribution of arrival time deviation. Suppose that we have Q classes and P_q be the set of patients in class q, $q = 1, \ldots, Q$. Without loss of generality, we can assume that patients within each P_q are numbered sequentially. Accordingly, we add the following symmetry breaking constraints to the TSM-PI and TSM-AO (and their SAA formulations):

$$x_{p,i} - \sum_{j=i+1}^{P} x_{p+1,j} \leq 0, \qquad \forall i = 1, \ldots, P,\ \forall p : p, p+1 \in P_q,\ q = 1,\ \ldots, Q, \tag{3.8}$$

indicating that, if patients p and $p+1$ are of the same class then p is scheduled before $p+1$.

3.5 Computational Results

In this section, we numerically study the approximate (statistical) relative gap ARG $:= \frac{\overline{v}^{\text{AO}}_{N'} - \overline{v}^{\text{PI}}_{N}}{\overline{v}^{\text{PI}}_{N}}$ between SOASP statistical upper $\overline{v}^{\text{AO}}_{N'}$ and lower $\overline{v}^{\text{PI}}_{N}$ bounds and the optimality of the appointment order policy under a range of parameter settings. In Section 3.5.1, we describe the set of SOASP instances that we use in our experiments and discuss other experimental setups. In Section 3.5.2, we obtain a near-optimal solution to the TSM-PI and TSM-AO via their SAA formulations (specifically, tight statistical lower $\overline{v}^{\text{PI}}_{N}$ and upper $\overline{v}^{\text{AO}}_{N'}$ bounds on the optimal values of these TSMs). In Section 3.5.3, we evaluate ARG between the statistical bounds on the optimal values of these TSMs, demonstrating the near-optimality of the appointment order policy in a wide range of SOASP parameter settings. We also identify parameter settings that result in a large gap and accordingly propose an alternative rescheduling policy that significantly shrinks the gap in Section 3.5.4. Finally, in Section 3.5.5, we demonstrate the benefits of incorporating uncertainty in

Table 3.3: Characteristics and parameters of the the SOASP instances.

Parameter	**Value/Distribution**
Number of Patients, P	12 patients
Patient mix	Hetero1: (9 return, 3 new), Hetero2: (6 return, 6 new), Homo1: (12 return), Homo2: (12 new)
Planned provider service hours, $[0, L]$	[0, expected duration for the patient mix]
Service time distributions	$d(\text{retrun}) \sim \text{LogN}(20, 16^2)$, $d(\text{new}) \sim \text{LogN}(30, 24^2)$
Arrival time deviation distributions	$N(\mu_u = -15, 0, \sigma_u^2 = 10^2, 20^2)$, $U[\underline{u} = -40, \overline{u} = 20, 40]$
Grace periods, G	10, 15, 20
Rescheduling Period, R	90 minutes

outpatient appointment scheduling.

3.5.1 Description of Experiments

We use assumptions and parameters settings made in prior outpatient appointment scheduling literature to construct 60 SOASP instances characterized by four patient mixes, five distributions of stochastic arrival time deviation, and three choices of the grace period. Each instance consists of twelve patients based on a typical outpatient scheduling problem studied in *Deceuninck et al.* (2018). Table 3.3 summarises the characteristics and parameters of these instances.

Heterogeneous instances (Hetero1 and Hetero2) consist of two types of patients; newly referred ("new") patients and followup ("return") patients (a typical patient mix in OPCs; see *Deceuninck et al.* (2018); *Zhu et al.* (2017) and references therein). Hetero1 is based on the problem studied in *Deceuninck et al.* (2018), where 75% of the patients are return and the remaining 25% are new patients. Hetero2 represents clinics where 50% of the patients are return and the remaining 50% are new patients. Homogenous instances (Homo1 and Homo2) represent clinics where patients are identical (i.e., of the same type).

The distributions for stochastic service duration of each patient type follow a lognormal distribution, $d(\text{type}) \sim \text{LogN}(\mu_d, \sigma_d^2)$ with mean μ_d and standard deviation σ_d as in *Deceuninck et al.* (2018). The lognormal distribution is a typical distribution for service duration in the appointment scheduling literature, and several empirical studies show how it accurately describes the shape of service duration distributions in a variety of service systems (see *Cayirli et al.* (2006); *Gul et al.* (2011); *Klassen and Rohleder* (1996); *Klassen and Yoogalingam* (2014) and references therein).

Consistent with the literature (see, e.g., *Cayirli et al.* (2006); *Deceuninck et al.* (2018); *Klassen and Yoogalingam* (2014)), we model the stochastic arrival time deviation u by: (1) normal distribution, $u \sim N(\mu_u, \sigma_u^2)$, with mean μ_u =-15 minutes, 0 minutes and standard deviation $\sigma_u = 10$ minutes, 20 minutes; and (2) uniform distribution $U[\underline{u}, \overline{u}]$ with two intervals $u \sim U[-40, 20]$ and $u \sim U[40, 40]$. We select the value of the grace period, G, from the set $\{10, 15, 20\}$ (e.g., G =15 minutes in *Deceuninck et al.* (2018)) and, unless stated otherwise, we consider a rescheduling period of 90 minutes in the TSM-PI (i.e., the model automatically decline a late patient arriving after $G + 90$ minutes past his/her scheduled time).

Finally, we consider two different cost structures for the objective function: (1) Cost1: $c^{\text{w}} = c^{\text{g}} = c^{\text{o}}$; and (2) Cost2: $c^{\text{w}} = 1, c^{\text{g}} = 5, c^{\text{o}} = 7.5$. For the first cost structure, each of the three objectives is equally important (this is a classical assumption in the domain of outpatient appointment scheduling, see, e.g., *Berg et al.* (2014); *Deceuninck et al.* (2018)). The second cost structure fixes the $c^{\text{o}}/c^{\text{g}}$ ratio to 1.5 as in *Deceuninck et al.* (2018), based on OPC practice (see *Cayirli et al.* (2006) and *Deceuninck et al.* (2018) for detailed discussions). We also adopt the classical assumption of identical waiting and idle time costs, i.e., $c^{\text{w}} = c_j^{\text{w}}$ and $c^{\text{g}} = c_j^{\text{g}}$ for all $j = 1, \ldots, P$ (see *Ahmadi-Javid et al.* (2017); *Deceuninck et al.* (2018) and the references therein).

We implemented our TSMs and the Monte Carlo Optimization approach using the AMPL2016 Programming language calling CPLEX V12.6.2 as a solver with default settings (our experiments showed no consistent benefits in any parameter tuning). We ran all experiments on an HP workstation running Windows Server 2012 with two Intel E5-2620-v4 processor, each with 8-Cores (16 total), 2.10GHz CPUs, and 128 GB shared RAM.

3.5.2 Obtaining Near-Optimal Solutions to the TSMs

For each SOASP instance and cost structure, we implemented the MCO algorithm in Section 3.4.1 with $N_o = 5$, $M = 20$, and $\epsilon = 0.1$ (i.e., Algorithm 3.1 terminates with N, $\overline{v}_N$ and $\overline{v}_{N'}$ whenever $|AOI_N| < 0.1$, see step 4).

Tables 3.6 and 3.7 in Appendix 3.7.4 present the approximate optimality index ($AOI_N = \frac{\overline{v}_{N'} - \overline{v}_N}{\overline{v}_{N'}}$) and the 95% confidence intervals (95%CI) of the statistical lower and upper bounds ($\overline{v}_N$ and $\overline{v}_{N'}$, receptively) on the objective values of TSM-PI and TSM-AO for each SOASP instance under Cost1 and Cost2, respectively, at the termination of Algorithm 3.1. In 3.7.5, we present and analyze the ranges of solution time of the TSM-PI and TSM-AO formulations for each SOASP instance. Herein we summarize the key findings of the MCO algorithm.

Clearly, $N = 100$ and $N = 200$ scenarios of service durations and arrival time deviations are sufficient to obtain a near-optimal solution to the TSM-PI and TSM-AO via their SAA formulations. First, $AOI_{100,200}$ ranges from 0.00 to 0.09 and from 0.00 to 0.06 for the TSM-PI and TSM-AO, respectively. Second, the 95%CI of $\overline{v}_N^{\text{PI, AO}}$ and $\overline{v}_{N'}^{\text{PI, AO}}$ with N=100, 200 are very tight (i.e., have a small variance). These results qualify $\overline{v}_N^{\text{PI}}$ and $\overline{v}_{N'}^{\text{AO}}$ as tight statistical estimates for the lower and upper bounds on the optimal value of each SOASP instance, respectively. In Section 3.5.3, we compute and analyze the the statistical relative gap (ARG $= \frac{\overline{v}_{N'}^{\text{AO}} - \overline{v}_N^{\text{PI}}}{\overline{v}_N^{\text{PI}}}$) between these two bounds.

Recall that during this experiment, we use $M = 20$ as the number of replication in the MCO Algorithm (see step 1). The tightness of the 95%CI of $\overline{v}_N$ and $\overline{v}_{N'}$ suggests that $M = 20$ replications are sufficient to get a tight confidence interval for both bounds. In other words, the standard error estimates are small enough so that (statistically) we have high confidence in our conclusions about these bounds.

3.5.3 Evaluating ARG and the Optimality of the AO Policy

Table 3.4 presents ARG$\% = \frac{\overline{v}^{\text{AO}}_{N'} - \overline{v}^{\text{PI}}_{N}}{\overline{v}^{\text{PI}}_{N}} \times 100\%$. Recall that ARG is a conservative approximation of the Gap $= \frac{v^{\text{AO}} - v^{\text{PI}}}{v^{\text{PI}}}$ between the two TSM approximations of SOASP. Hence, small ARG$\%$ indicates that the AO policy is near-optimal. We observe the following about ARG$\%$ and the optimality of the AO policy.

First, the AO policy is near-optimal in a wide range of parameter settings. Specifically, the ARG$\%$ is small if (i) the patients arrive on time or early on average with a small variability in arrival time or (ii) the grace period is relatively long. For example, ARG$\%$ ranges from 0.1% to 4% when $u \sim N(-15, 10^2)$ and a majority of ARG$\%$ are less than 7% when G = 15, 20. This makes sense because under these parameter settings, the majority of the patients are on-time or early (i.e., arrive within interval E or G) and so should be served in their scheduled order (i.e., according to AO).

Second, the AO policy becomes sub-optimal under high variability in arrival time and short grace periods. For example, with $G = 10$ minutes, increasing the variability from $u \sim N(-15, 10^2)$ to $u \sim N(-15, 20^2)$ increased the range of ARG$\%$ from 1–4% to 10–19%. In addition, ARG$\%$ is significantly larger when the patients are likely to arrive after their grace periods (e.g., with G = 10 and $u \sim U[-40, 40]$, $\mathbb{P}(u > G) \approx 0.38$; see columns 10–11 in Table 4). This is because, in this case, the AO policy declines all patients arriving in interval R or DE, which yields unnecessary provider idling. In contrast, TSM-PI may reschedule some those arriving within interval R.

Third, in all parameter settings, the AO policy becomes closer to optimal (i.e., ARG$\%$ decreases) as the grace period lengthens. This is particularly significant when there is a high degree in patient lateness (e.g., when $u \sim U[40, 40]$) and when idling is costly (see, e.g., the last column of Table 3.4). Intuitively, the longer the grace period, the lower the probability of arriving beyond the grace period (i.e., within interval R or DE) and thus the less the number of patients to be rescheduled. For example, with u$\sim$ $U[40, 40]$ and as G increases from 10 to 20, the probabilities of arriving beyond the grace period decreases from $\mathbb{P}(u > 10) \approx 0.38$ to $\mathbb{P}(u > 20) = 0.25$.

3.5.4 Alternative Rescheduling Policy

We propose and evaluate the performance of an alternative rescheduling policy, which we call Neighbor Swapping (NS), for those parameter settings under which the AO policy is suboptimal. The NS policy has two schemes; a priority queueing scheme (in Algorithm 3.2) and a serving scheme (in Algorithm 3.3). A patient receptionist can implement the NS policy as follows: (i) using the priority queueing scheme, s/he maintains a priority queue of patients waiting for service and updates the priority whenever a patient arrives, and (ii) using the serving scheme, s/he decides whether the provider, whenever idle, should stay idle or start serving the patient with the highest priority in the queue.

Note that the NS policy respects the appointment order of early and punctual arrivals and deprioritizes or even declines late patients. These two properties are particularly demanded in both the OPC practice and the appointment scheduling literature (see, e.g., *Deceuninck et al.* (2018); *Glowacka et al.* (2017) and the references therein).

We evaluate the performance of the NS policy for those parameters settings under which the AO policy is sub-optimal (i.e., SOASP instances with large ARG$\%$ values, as marked in bold in Table 3.4). First, we fix the initial appointment sequence and the corresponding scheduled arrival time to the optimal ones to the TSM-AO. Second, for each $R \in \{5, 10, 0.5G, G, 2G, 3G\}$ we simulate appointment system under the NS policy for N_{NS}=10,000 scenarios of service durations and arrival time deviations (d_p^n, u_p^n) for all $p = 1, \dots, P$ and $n = 1, \dots, N_{\text{NS}}$. For each each $n = 1, \dots, N_{\text{NS}}$, we compute w^n, g^n, w_{P+1}^n, and $v^{\text{NS}} = \frac{1}{N_{\text{NS}}} \sum_{n=1}^{N_{\text{NS}}} \sum_{j=1}^{P} \left(c_j^{\text{w}} w_j^n + c_j^{\text{g}} g_j^n \right) + c^{\text{o}} w_{P+1}^n$. We repeat this simulation 20 times, each time with a new sample of N_{NS} scenarios. Finally, we take the average of the 20 objective function values to obtain $\overline{v}^{\text{NS}}$.

Table 3.4: ARG% for each SOASP instance under two cost structures: (1) Cost1: $c^{\text{w}} = c^{\text{g}} = c^{\text{o}}$; and (2) Cost2: $c^{\text{w}} = 1, c^{\text{g}} = 5, c^{\text{o}} = 7.5$. Results are based on the average across 20 random instances for each combination of patient mix, stochastic arrival distribution, grace period and cost structures. ARG% values greater than or equal to 10% are marked in bold.

Hetero1

G	N(0,10^2)		N(-15,10^2)		N(-15,20^2)		U[-40,20]		U[-40,40]	
	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2
10	**18%**	**17%**	4%	1%	**15%**	**12%**	**14%**	**29%**	**40%**	**65%**
15	4%	2%	3%	1%	7%	4%	7%	**11%**	**24%**	**49%**
20	0.3%	0.1%	2%	0.3%	0.3%	2%	0.5%	3%	**24%**	**29%**

Hetero2

G	N(0,10^2)		N(-15,10^2)		N(-15,20^2)		U[-40,20]		U[-40,40]	
	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2
10	**15%**	**15%**	3%	1%	**14%**	**12%**	**16%**	**30%**	**27%**	**67%**
15	5%	7%	4%	0.1%	6%	4%	7%	**13%**	**28%**	**53%**
20	0.3%	3%	1%	0.1 %	3%	1%	0.3	4%	**26%**	**23%**

Homo1

G	N(0,10^2)		N(-15,10^2)		N(-15,20^2)		U[-40,20]		U[-40,40]	
	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2
10	**10%**	**20%**	1%	1%	**12%**	**19%**	**14%**	**21%**	**29%**	**68%**
15	2%	6%	1%	0.1 %	3%	4%	6%	6%	**26%**	**52%**
20	0.4%	1%	0.8%	0.1%	1%	2%	0.3%	0.2%	**24%**	**29%**

Homo2

G	N(0,10^2)		N(-15,10^2)		N(-15,20^2)		U[-40,20]		U[-40,40]	
	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2
10	**12%**	**18%**	1%	1%	**10%**	**18%**	**13%**	**20%**	**24%**	**73%**
15	3%	3%	0.1%	1%	3%	4%	6%	7%	**27%**	**54%**
20	0.2%	0.6%	0.1%	0.3%	3%	2%	0.3%	0.7%	**25%**	**34%**

Algorithm 3.2: Neighbor Swapping Policy–Priority Queueing Scheme

```
Input: scheduled arrival times $\{t_i\}_{i=1}^{P}$, grace period ($G$), rescheduling period ($R$)
for (all patients arriving at the clinic) do
    Denote the set of the initial positions of patients currently waiting as $\mathcal{W}$.
    Denote the initial position of the current arrival (patient) as $i$.
    if $a_i > t_i + G + R$ then    /* if the patient $i$ arrives in interval DE. */
        Decline this patient.
    else
        if $a_i \leq t_i + G$ then    /* if the patient $i$ arrives in interval E or G. */
            This patient $i$ has lower priority than all patients in $\{j \in \mathcal{W} : j \leq i-1\}$, and higher
            priority than all patients in $\{j \in \mathcal{W} : j \geq i+1\}$.    /* respects the
            appointment order of on-time patients. */
        end
        if $t_i + G < a_i \leq t_i + G + R$ and $a_i \leq a_{i+1}$ then    /* if the patient $i$ arrives
        in interval R and earlier than patient $(i+1)$. */
            This patient $i$ has lower priority than all patients in $\{j \in \mathcal{W} : j \leq i-1\}$, and higher
            priority than all patients in $\{j \in \mathcal{W} : j \geq i+1\}$. /* keep the order of $i$ and
            $i+1$. */
        end
        if $t_i + G < a_i \leq t_i + G + R$ and $a_i > a_{i+1}$ then    /* if the patient $i$ arrives
        in interval R and later than patient $(i+1)$. */
            This patient $i$ has lower priority than all patients in $\{j \in \mathcal{W} : j \leq i-1\} \cup \{i+1\}$, and
            higher priority than all patients in $\{j \in \mathcal{W} : j \geq i+2\}$.  /* swap the order of
            $i$ and $i+1$. */
        end
    end
end
```

Note that the NS policy is a feasible rescheduling policy to SOASP. Therefore, $\overline{v}^{\text{NS}}$ serves as a statistical upper bound on the optimal value of SOASP. Accordingly, we compute ARG$\%$ under NS as $\frac{\overline{v}^{\text{NS}}-\overline{v}_N^{\text{PI}}}{\overline{v}_N^{\text{PI}}}$.

For those parameter settings under which the AO policy is sub-optimal (i.e., SOASP instances with large ARG$\%$ values, as marked in bold in Table 3.4), Figures 3.2–3.3 compare ARG$\%$ under the AO and NS policies for the Hetero and Homo instances, respectively. In addition, Table 3.9 in Appendix 3.7.6 reports the best selection of rescheduling period R^* from $\{5, 10, 0.5G, G, 2G, 3G\}$ associated with the minimum (i.e., tightest) $\overline{v}^{\text{NS}}$ in each parameter setting. We observe from Fig-

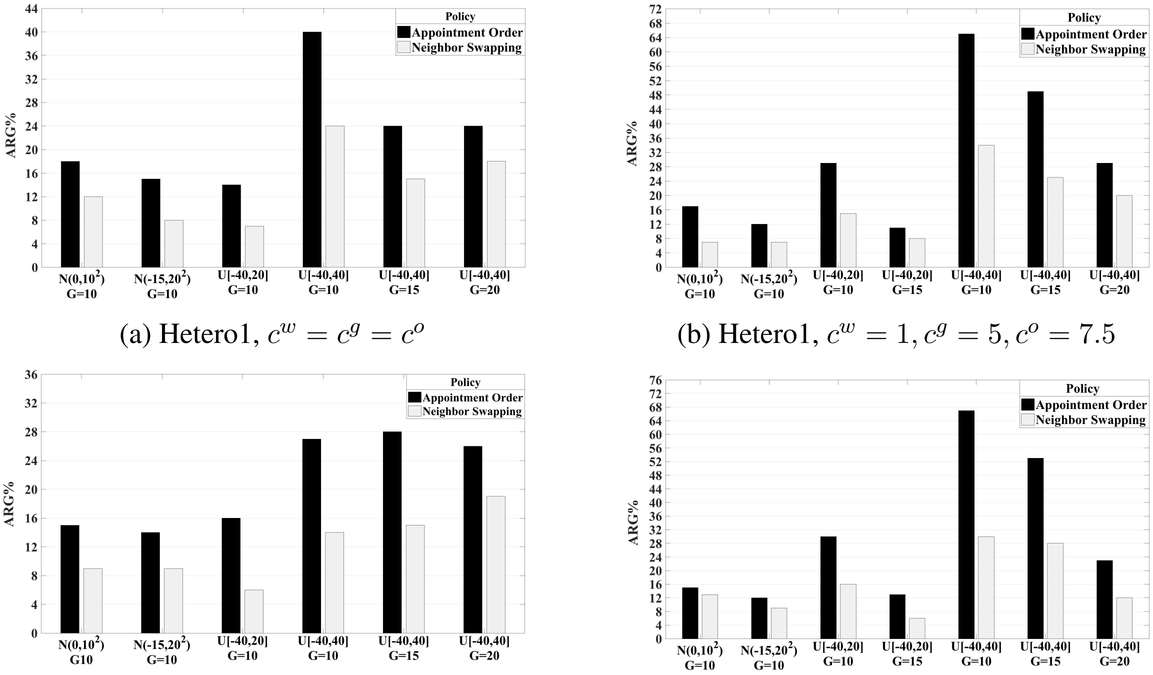
Algorithm 3.3: Neighbor Swapping Policy–Serving Scheme

```
Input: scheduled arrival times $\{t_i\}_{i=1}^{P}$, grace period ($G$), rescheduling period ($R$)
while (provider is idle) do
    Denote the set of the initial positions of patients currently waiting as $\mathcal{W}$.
    Denote $i^*$ as the initial position of the patient with the highest priority in $\mathcal{W}$.
    Denote the current clock time as $T$.
    if $\mathcal{W} = \emptyset$ then                /* no patients waiting for service. */
        The provider should wait.
    else
        if $\mathcal{W} \neq \emptyset$ and $T \geq t_j + G$ for all $j = 1, \ldots, (i^* - 1)$ then      /* all patients $j$,
        with $j \leq i^* - 1$, either are late or have been served. */
            The provider should start serving patient $i^*$. BREAK.
        else
            The provide should wait.
        end
    end
end
```

ures 3.2–3.3 that the NS policy significantly and consistently shrinks the ARG% in all challenging parameter settings (i.e., those with high variability in arrival time and short grace periods). For example the NS policy is near-optimal when $G = 10$ and $u \sim \{N(0, 10^2),\ N(-15, 20^2),\ U[-40, 20]\}$ with ARG% ranging from 2% to 16%. In addition, when $u \sim U[-40, 40]$, NS cuts the ARG% by half on average from those obtained under AO. This indicates that the NS policy can effectively reschedule the appointments and avoid unnecessary idling.

As shown in Table 3.11 in 3.7.7, the average number of declined appointments per day under NS is approximately zero under $G = 10$ and $u \sim \{N(0, 10^2),\ N(-15, 20^2),\ U[-40, 20]\}$, and is significantly less than under the AO policy when $u \sim U[-40, 40]$. This is another important property of the NS policy because a declined patient will have to be rescheduled for another day, which is inconvenient to the patient and has a cost to the system.

Finally, it is noteworthy to mention that we repeat the same simulation using the optimal schedule to the TSM-PI as the initial schedule. The obtained results on the NS policy are similar. Additionally, we use the same simulation steps to evaluate the performance of the NS policy for those parameter settings under which AO is near-optimal (i.e., SOASP instances with small ARG% val-



(a) Hetero1, $c^w = c^g = c^o$

(b) Hetero1, $c^w = 1, c^g = 5, c^o = 7.5$

(c) Hetero2, $c^w = c^g = c^o$

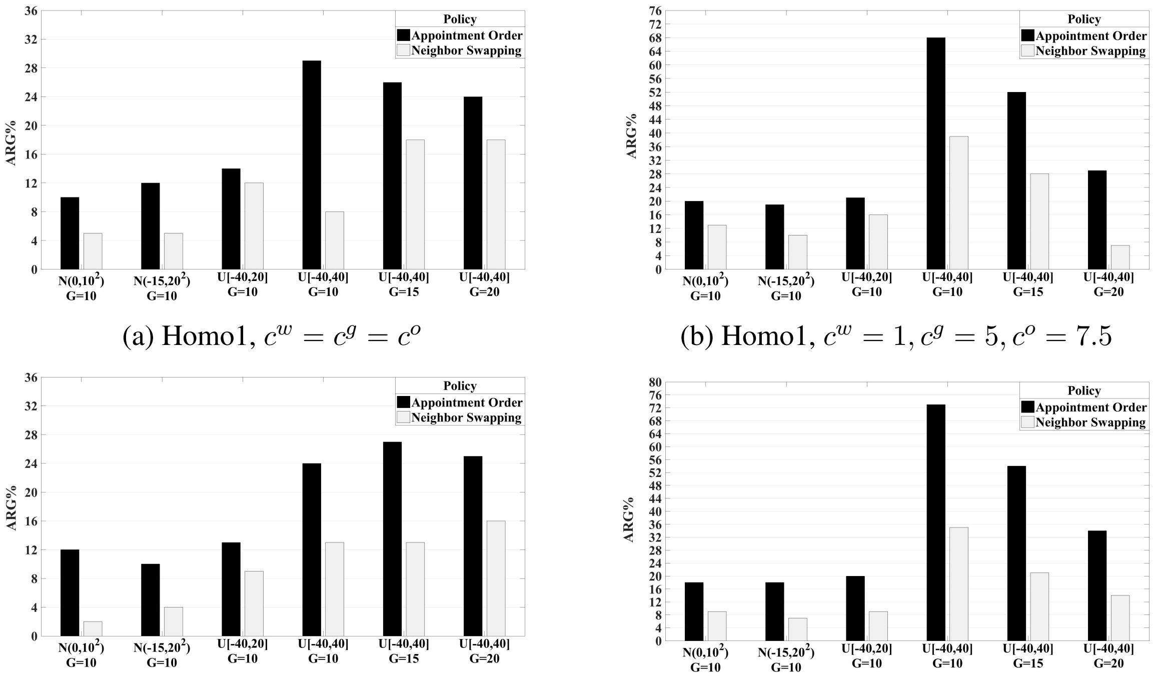
(d) Hetero2, $c^w = 1, c^g = 5, c^o = 7.5$

Figure 3.2: Comparisons of the ARG$\%$ values under appointment order and neighbor swapping policies for the Hetero instances.

ues in Table 3.4). The obtained ARG% values under NS are very similar to ARG% values obtained under AO, demonstrating the near-optimality of NS under those parameter settings. This makes sense because under these settings the majority of patients are punctual or early, and so should be served in the appointment order according to both the priority queuing and serving schemes.

3.5.5 Value of Modeling Stochastic Arrivals

In this section, we demonstrate the benefit of modeling stochastic arrivals in outpatient appointment scheduling. For each SOASP instance, we solve the TSM-AO twice, with one assuming punctual arrivals (i.e., zero arrival time deviations) and the other considering stochastic arrivals. We denote the obtained schedules as S^0 and S^u, respectively. We then evaluate the cost $E[Q^{AO}(x, t, \xi)]$ of S^0, denoted as C^0, by reevaluating the schedule S^0 under a set of 10,000 scenarios with stochastic



(a) Homo1, $c^w = c^g = c^o$

(b) Homo1, $c^w = 1, c^g = 5, c^o = 7.5$

(c) Homo2, $c^w = c^g = c^o$

(d) Homo2, $c^w = 1, c^g = 5, c^o = 7.5$

Figure 3.3: Comparisons of the ARG% values under appointment order and neighbor swapping policies for the Homo instances.

arrivals, while C^{u} denotes the corresponding cost of S^{u}. We compute the relative increase in cost $\text{RC}^{\text{(punc)}} := \frac{C^0 - C^{\text{u}}}{C^{\text{u}}} \times 100\%$.

We also compare S^{u} with the well-known and widely employed deterministic schedule S^{m} that assigns appointment slot to each patient by the mean service duration of his/her type. For all heterogeneous instances, we use the optimal sequence to the TSM-AO to assign patient types to appointments in S^{m} and fix the length of each appointment slot to the mean service duration of the assigned type to it. We then obtain the cost C^{m} of ignoring uncertainty by reevaluating S^{m} under a set of 10,000 scenarios with stochastic arrivals and service durations. We compute the relative increase in cost $\text{RC}^{\text{(mean)}} := \frac{C^{\text{m}} - C^{\text{u}}}{C^{\text{u}}} \times 100\%$.

Table 3.5 presents $\text{RC}^{\text{(punc)}}$ and $\text{RC}^{\text{(mean)}}$, including the average and maximum among all combinations of patient mix and distribution of stochastic arrival time deviation. From this table, we observe that the cost of S^{u} (i.e., considering stochastic arrivals) is significantly lower than that of

Table 3.5: The average and maximum relative cost gaps $\mathrm{RC}^{(\mathrm{punc})}$ and $\mathrm{RC}^{(\mathrm{mean})}$ as functions of patient mix and stochastic arrival distributions. Results are based on 20 random instances for each parameter setting.

				Hetero1						
	N(0,10^2)		**N(-15,10^2)**		**N(-15,20^2)**		**U[-40,20]**		**U[-40,40]**	
	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2
$\mathrm{RC}^{(\mathrm{punc})}$	18–31	7–18	10–24	1–14	17–26	6–20	7–20	7–18	26–28	40–46
$\mathrm{RC}^{(\mathrm{mean})}$	24–41	13–38	31–54	8–21	28–47	12–27	24–39	18–29	27–37	44–55
				Hetero2						
	N(0,10^2)		**N(-15,10^2)**		**N(-15,20^2)**		**U[-40,20]**		**U[-40,40]**	
	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2
$\mathrm{RC}^{(\mathrm{punc})}$	11–20	8–21	7–15	6–22	9–19	7–22	9–18	10–20	38–46	45–52
$\mathrm{RC}^{(\mathrm{mean})}$	16–20	15–34	25–42	8–25	17–32	14–31	18–34	22–32	34–47	70–78
				Homo1						
	N(0,10^2)		**N(-15,10^2)**		**N(-15,20^2)**		**U[-40,20]**		**U[-40,40]**	
	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2
$\mathrm{RC}^{(\mathrm{punc})}$	15–17	3–17	7–30	3–15	7–15	6–19	7–18	7–17	26–35	29–42
$\mathrm{RC}^{(\mathrm{mean})}$	19–28	15–32	25–59	8–22	19–34	12–25	13–28	15–26	13–32	23–40
				Homo2						
	N(0,10^2)		**N(-15,10^2)**		**N(-15,20^2)**		**U[-40,20]**		**U[-40,40]**	
	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2	Cost1	Cost2
$\mathrm{RC}^{(\mathrm{punc})}$	13–16	10–21	9–27	7–22	10–23	8–20	8–18	10–32	28–32	27–40
$\mathrm{RC}^{(\mathrm{mean})}$	23–33	21–33	22–50	10–26	22–42	18–30	16–33	18–39	21–35	49–64

S^0 (i.e., ignoring stochastic arrivals). Overall, the average and maximum $\mathrm{RC}^{(\mathrm{punc})}$ range in 7%–45% and 14%–52%, respectively (see Tables 3.11–3.12 in Appendix 3.7.8 for the improvement in scheduling metric). In addition, we observe that the cost of S^{u} (i.e., considering stochasticity) is significantly lower than that of S^{m} (i.e., ignoring stochasticity). The average and maximum $\mathrm{RC}^{(\mathrm{mean})}$ range in 8%–70% and 20%–78%.

3.6 Conclusion and Chapter Summary

In this chapter, we studied SOASP for OPC scheduling under stochastic arrival times and service durations. We consider for an OPC manager who needs to design an appointment schedule and a rescheduling policy for a single provider and a set of patients, where each patient has a known probability distribution of arrival time deviations and service durations. The objective is to minimize the expected total cost of patient waiting time, provider idle time, and provider overtime.

By deriving two-stage approximations of SOASP and testing them in extensive numerical experiments, we show that the AO policy is near-optimal in a wide range of realistic parameter settings. We also identify parameter settings that result in sub-optimality of the AO policy. Accordingly, we propose an alternative policy based on neighbor-swapping that leads to a significantly better performance.

To the best of our knowledge, and according to the recent review of outpatient appointment systems by *Ahmadi-Javid et al.* (2017), this chapter presents the first stochastic programming approach to SOASP that considers (1) patient heterogeneity, (2) optimizing both the initial appointment sequencing and scheduling decisions, and (3) the possibility of rescheduling (i.e., resequencing or declining).

3.7 Appendix

3.7.1 Proof of Proposition 3.4.1

Proposition 3.4.1 *$t_i^{min} \leq t_i \leq t_i^{max}$, where $t_i^{min} = (i-1)G$ and $t_i^{max} = L - (P-i)G$ for all $i \in [P]$.*

Proof. Recall that in Assumption A5 we assume that the scheduled times of two consecutive patients are separated by at least one grace period. This is ensured by constraints (3.1e). Given that $t_i \leq L$ by (3.1e), then the following bounds on the scheduled time are valid:

$$(i-1)G \leq t_i \leq L - (P-i)G, \qquad \forall i \in [P] \tag{3.9}$$

Therefore, we set $t_i^{\text{min}} = (i-1)G$ and $t_i^{\text{max}} = L - (P-i)G$ in (3.6a)–(3.6b).

□

3.7.2 SMILP Reformulation of the SAA Formulation (3.4)

$$v_N^{\text{PI}} = \underset{a,s,w,g}{\text{minimize}} \ \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{P}\left(c_j^{\text{w}} w_j^n + c_j^{\text{g}} g_j^n\right) + c^{\text{o}} w_{P+1}^n \tag{3.10a}$$

Subject to: (3.1b) – (3.1f)

$$\begin{aligned}
&(3.2\text{b}) - (3.2\text{g}) && \forall n\\
&(3.3\text{d}) - (3.3\text{k}) && \forall n \in \mathcal{N}^{\text{EG}}\\
&z_j^n = 0, && \forall j \in [P],\ \forall n \in \mathcal{N}^{\text{EG}}\\
&y_{j,j}^n = 1, && \forall j \in [P],\ n \in \mathcal{N}^{\text{EG}}\\
&\widetilde{d}_j = \sum_{p=1}^{P}\sum_{i=1}^{P} d_p^n \beta_{p,i,j}^n, && \forall j \in [P],\ \forall n \notin \mathcal{N}^{\text{EG}}\\
&a_j^n = \sum_{i=1}^{P}\left(\alpha_{i,j}^n + \sum_{p=1}^{P} u_p^n \beta_{p,i,j}^n\right), && \forall j \in [P], \forall n \notin \mathcal{N}^{\text{EG}}\\
&(3.2\text{j}) - (3.2\text{k}) && \forall n \notin \mathcal{N}^{\text{EG}}\\
&w_j^n \geq s_j^n - a_j^n - M_j^n\left(z_j^n + \sum_{p=1}^{P}\sum_{i=1}^{P} e_p^n \beta_{p,i,j}^n\right), && \forall j \in [P], \forall n \notin \mathcal{N}^{\text{EG}}\\
&w_j^n \geq s_j - \sum_{i=1}^{P}\alpha_{i,j}^n - M_j^n\left(2 - (1 - z_j^n) - \sum_{p=1}^{P}\sum_{i=1}^{P} e_p^n \beta_{p,i,j}^n\right), && \forall j \in [P],\ \forall n \notin \mathcal{N}^{\text{EG}}\\
&(3.2\text{n}) - (3.2\text{p}) && \forall n \notin \mathcal{N}^{\text{EG}}\\
&(3.2\text{q}) - (3.2\text{r}) && \forall n
\end{aligned}$$

3.7.3 Proof of Proposition 3.4.3

Proposition 3.4.3 $M_j^n = L + \max_{p=1,\ldots,P}\{|u_p^n|\} + (j-1)\max_{p=1,\ldots,P}\{d_p^n\}$ *for* $j = 1, \ldots, P$ *and* $n \in [N]$ *is a valid value for the Big-M constant in constraints* (3.2l)–(3.2m) *and* (3.3g)–(3.3h).

Proof. We prove the validity of M_j in constraints (3.2l)–(3.2m) of the TSM-PI and the proof of the validity of M_j in constraints (3.3g)–(3.3h) of the TSM-AO is similar. For notational convenience, we suppress the scenario index n from the scenario-dependent parameters, variables, and constraints.

From constraints (3.2l)–(3.2m) and (3.2r) we have

$$\begin{aligned} w_j &= \max\{s_j - \max\{a_j, t_j\}, 0\} \\ &= \max\left\{s_j - \max\{t_j + u_j, t_j\}, 0\right\}, \quad \text{for } j = 2, \ldots P \\ \therefore\ w_j^{\text{max}} &= \left\{s_j^{\text{max}} - t_j^{\text{min}}, 0\right\} \end{aligned} \tag{3.11}$$

where $u_j \equiv \sum\limits_{i=1}^{p} \sum_{p=1}^{P} u_p x_{p,i} y_{i,j}$, $t_j \equiv \sum\limits_{i=1}^{P} t_i y_{i,j}$, w_j^{max} is the maximum waiting time of appointment j, and s_j^{max} is the maximum actual start time. It follows from (3.11) that to preserve optimality, M_j should be greater than or equal to w_j^{max}, i.e., $M_j \geq w_j^{\text{max}}$. Next, we derive an upper bound on s_j^{max} in (3.11). By constraints (3.2j)–(3.2k), we can compute the value of s_j recursively as follow:

$$\begin{aligned} &s_1 = \max\{t_1, t_1 + u_1\} \\ &s_2 = \max\{s_1 + \widetilde{d}_1, a_2\} = \max\{t_1 + \widetilde{d}_1, t_1 + u_1 + \widetilde{d}_1, t_2 + u_2\} \\ &s_3 = \max\{s_2 + \widetilde{d}_2, t_3 + u_3\} = \max\{t_1 + \widetilde{d}_1 + \widetilde{d}_2, t_1 + u_1 + \widetilde{d}_1 + \widetilde{d}_2, t_2 + u_2 + \widetilde{d}_2, t_3 + u_3\} \\ &\ldots \\ &s_j = \max\left\{t_1 + \sum_{k=1}^{j-1} \widetilde{d}_k, \max_{i=1,\ldots,j} \{u_i + t_i + \sum_{k=i}^{j-1} \widetilde{d}_k\}\right\} \\ &s_j^{\text{max}} \leq L + \max_{p=1,\ldots,P}\{|u_p^n|\} + (j-1) \max_{p=1,\ldots,P}\{d_p^n\} \end{aligned} \tag{3.12}$$

where $|C|$ denote the absolute value of C.The last inequality (3.12) holds because (1) $t_i \leq L$ for all i by constraints (3.1d), (2) $u_i \leq \max\limits_{p=1,\ldots,P} |u_p|$ for all i, and (3) $\widetilde{d}_j \leq \max\limits_{p=1,\ldots,P} d_p$. It follows from (3.12) that

$$w_j^{\max} \leq L + \max_{p=1,\ldots,P}\{|u_p^n|\} + (j-1)\max_{p=1,\ldots,P}\{d_p^n\} = M_j$$

□

3.7.4 Convergence Results at the Selected Sample Size

Table 3.6: The Approximate Optimality Index (AOI_N) between the statistical lower bound $\overline{v}_N$ and upper bound $\overline{v}_{N'}$ on the objective values of TSM-PI and TSM-AO and their 95% Confidence Interval (95%CI) at the selected sample size N for each SOASP instance under Cost1 ($c^w = c^g = c^o$).

Inst	Dist	G	Size, N	95%CI$\overline{v}_N$	95%CI$\overline{v}_{N'}$	AOI_N	Size, N	95%CI$\overline{v}_N$	95%CI$\overline{v}_{N'}$	AOI_N
			(TSM-PI)				(TSM-AO)			
Hetero1	$N(0, 10^2)$	**10**	100	[200, 203]	[217, 220]	0.07	100	[228, 232]	[236, 237]	0.02
		15	100	[257, 261]	[282, 284]	0.07	100	[250, 255]	[268, 271]	0.05
		20	200	[285, 289]	[301, 303]	0.05	200	[279, 281]	[286, 288]	0.02
	$N(-15, 10^2)$	**10**	200	[223, 225]	[231, 233]	0.03	200	[224, 226]	[230, 234]	0.03
		15	200	[223, 227]	[230, 231]	0.02	200	[224, 227]	[231, 232]	0.03
		20	200	[225, 228]	[230, 232]	0.02	200	[224, 226]	[230, 232]	0.03
	$N(-15, 20^2)$	**10**	100	[185, 189]	[197, 198]	0.05	200	[203, 205]	[214, 216]	0.05
		15	200	[225, 229]	[238, 242]	0.05	200	[230, 233]	[229, 232]	0.01
		20	200	[247, 250]	[269, 272]	0.08	200	[243, 245]	[248, 249]	0.02
	$U[-40, 20]$	**10**	200	[182, 182]	[185, 187]	0.03	200	[211, 213]	[198, 202]	0.06
		15	200	[227, 229]	[236, 237]	0.03	200	[239, 243]	[244, 246]	0.01
		20	200	[280, 283]	[285, 287]	0.01	200	[277, 280]	[282, 284]	0.02
	$U[-40, 40]$	**10**	100	[127, 130]	[129, 131]	0.01	200	[181, 182]	[182, 183]	0.01
		15	100	[139, 154]	[159, 161]	0.06	200	[199, 201]	[201, 202]	0.01
		20	100	[177, 184]	[190, 193]	0.05	200	[221, 223]	[226, 226]	0.02
Hetero2	$N(0, 10^2)$	**10**	200	[227, 230]	[250, 253]	0.08	200	[252, 257]	[265, 266]	0.03
		15	200	[276, 279]	[292, 294]	0.05	200	[293, 296]	[291, 293]	0.01
		20	200	[311, 315]	[314, 316]	0.01	200	[309, 312]	[312, 314]	0.01
	$N(-15, 10^2)$	**10**	200	[251, 254]	[259, 261]	0.03	200	[254, 258]	[259, 261]	0.01
		15	200	[254, 257]	[259, 261]	0.02	200	[254, 257]	[258, 260]	0.01
		20	200	[255, 258]	[260, 261]	0.01	200	[255, 258]	[258, 260]	0.01
	$N(-15, 20^2)$	**10**	100	[210, 213]	[227, 228]	0.07	200	[232, 236]	[240, 241]	0.02
		15	200	[243, 247]	[249, 251]	0.02	200	[255, 259]	[256, 259]	0.01
		20	200	[266, 270]	[272, 274]	0.02	200	[269, 270]	[274, 276]	0.02
	$U[-40, 20]$	**10**	200	[202, 202]	[208, 211]	0.04	200	[234, 236]	[233, 235]	0.01
		15	200	[252, 255]	[260, 261]	0.03	200	[264, 267]	[269, 271]	0.02
		20	200	[309, 313]	[313, 314]	0.01	200	[306, 309]	[310, 311]	0.01
	$U[-40, 40]$	**10**	100	[159, 163]	[181, 183]	0.09	100	[190, 200]	[204, 206]	0.03
		15	200	[171, 175]	[180, 184]	0.05	200	[221, 222]	[221, 223]	0.01
		20	200	[196, 199]	[217, 220]	0.09	200	[238, 240]	[250, 250]	0.04
Homo1	$N(0, 10^2)$	**10**	200	[203, 207]	[214, 216]	0.04	200	[214, 224]	[224, 227]	0.02
		15	200	[247, 251]	[253, 254]	0.02	200	[256, 258]	[252, 254]	0.02
		20	200	[277, 282]	[280, 282]	0.01	**100**	[277, 279]	[280, 282]	0.01
	$N(-15, 10^2)$	**10**	200	[219, 221]	[222, 224]	0.02	200	[219, 221]	[219, 221]	0.01
		15	200	[219, 222]	[223, 225]	0.02	200	[216, 221]	[220, 222]	0.01
		20	200	[222, 225]	[226, 227]	0.01	200	[221, 226]	[225, 226]	0.01
	$N(-15, 20^2)$	**10**	200	[173, 176]	[192, 193]	0.09	200	[194, 196]	[206, 207]	0.02
		15	200	[211, 215]	[217, 219]	0.02	200	[223, 225]	[219, 221]	0.02
		20	200	[238, 242]	[243, 244]	0.01	200	[242, 244]	[242, 244]	0.01
	$U[-40, 20]$	**10**	200	[174, 175]	[176, 179]	0.02	200	[203, 204]	[201, 203]	0.01
		15	200	[220, 223]	[228, 229]	0.03	200	[232, 235]	[236, 238]	0.01
		20	200	[281, 284]	[286, 287]	0.01	200	[279, 282]	[283, 285]	0.01
	$U[-40, 40]$	**10**	200	[197, 202]	[199, 206]	0.02	200	[244, 249]	[258, 260]	0.04
		15	200	[180, 182]	[178, 181]	0.01	200	[187, 193]	[192, 194]	0.01
		20	200	[188, 192]	[192, 196]	0.02	200	[215, 217]	[219, 220]	0.01
Homo2	$N(0, 10^2)$	**10**	200	[294, 300]	[311, 314]	0.04	200	[322, 329]	[333, 335]	0.02
		15	200	[354, 360]	[364, 367]	0.02	200	[370, 373]	[365, 368]	0.01
		20	200	[389, 395]	[395, 397]	0.01	200	[386, 390]	[392, 395]	0.01
	$N(-15, 10^2)$	**10**	200	[337, 340]	[343, 346]	0.02	200	[332, 335]	[338, 340]	0.02
		15	200	[337, 341]	[343, 346]	0.01	200	[334, 337]	[340, 343]	0.02
		20	200	[337, 341]	[343, 346]	0.01	200	[292, 329]	[340, 342]	0.06
	$N(-15, 20^2)$	**10**	200	[277, 284]	[301, 303]	0.06	200	[294, 297]	[310, 313]	0.05
		15	200	[311, 317]	[323, 326]	0.03	200	[329, 333]	[324, 327]	0.02
		20	200	[338, 343]	[348, 351]	0.02	200	[345, 347]	[348, 351]	0.01
	$U[-40, 20]$	**10**	200	[266, 270]	[262, 264]	0.02	200	[290, 294]	[303, 304]	0.04
		15	200	[318, 321]	[329, 332]	0.03	200	[334, 339]	[341, 343]	0.04
		20	200	[386, 391]	[394, 397]	0.02	200	[381, 385]	[389, 391]	0.04
	$U[-40, 40]$	10	**100**	[208, 218]	[215, 219]	0.01	200	[254, 258]	[255, 256]	0.01
		15	**100**	[210, 217]	[211, 220]	0.01	**100**	[269, 276]	[274, 278]	0.01
		20	**100**	[237, 246]	[249, 253]	0.03	200	[298, 300]	[304, 305]	0.02

Table 3.7: The Approximate Optimality Index (AOI_N) between the statistical lower bound $\overline{v}_N$ and upper bound $\overline{v}_{N'}$ on the objective values of TSM-PI and TSM-AO and their 95% Confidence Interval (95%CI) at the selected sample size N for each SOASP instance under Cost2 ($c^w = 1, c^g = 5, c^o = 7.5$).

Inst	Dist	G	Size, N	95%CI$^{\overline{v}}N$	95%CI$^{\overline{v}}N'$	AOI_N	Size, N	95%CI$^{\overline{v}}N$	95%CI$^{\overline{v}}N'$	AOI_N
			(TSM-PI)				(TSM-AO)			
Hetero1	$N(0, 10^2)$	10	200	[470, 481]	[481, 488]	0.02	200	[512, 523]	[527, 528]	0.01
		15	200	[596, 609]	[598, 603]	0.06	200	[622, 631]	[615, 620]	0.02
		20	200	[723, 736]	[738, 741]	0.01	200	[722, 728]	[729, 733]	0.01
	$N(-15, 10^2)$	10	200	[554, 564]	[569, 575]	0.02	200	[547, 554]	[561, 565]	0.02
		15	200	[565, 575]	[572, 578]	0.01	200	[562, 572]	[570, 574]	0.01
		20	200	[580, 590]	[589, 593]	0.01	200	[580, 590]	[583, 587]	0.01
	$N(-15, 20^2)$	10	200	[447, 459]	[459, 464]	0.01	200	[486, 492]	[509, 513]	0.04
		15	200	[524, 537]	[534, 538]	0.01	200	[560, 569]	[553, 559]	0.02
		20	200	[620, 631]	[621, 625]	0.02	200	[637, 642]	[636, 642]	0.02
	$U[-40, 20]$	10	100	[407, 410]	[379, 387]	0.05	200	[464, 471]	[494, 497]	0.05
		15	100	[524, 533]	[553, 555]	0.04	100	[567, 579]	[594, 598]	0.03
		20	100	[720, 726]	[741, 744]	0.02	100	[696, 702]	[739, 742]	0.05
	$U[-40, 40]$	10	100	[241, 251]	[244, 250]	0.00	200	[409, 411]	[407, 409]	0.00
		15	100	[314, 324]	[329, 336]	0.04	200	[476, 480]	[473, 475]	0.01
		20	200	[483, 498]	[498, 504]	0.02	200	[631, 638]	[639, 640]	0.01
Hetero2	$N(0, 10^2)$	10	200	[490, 502]	[494, 500]	0.00	200	[563, 574]	[575, 577]	0.01
		15	100	[627, 635]	[656, 660]	0.04	100	[656, 671]	[678, 680]	0.02
		20	200	[619, 629]	[629, 633]	0.01	200	[605, 613]	[622, 626]	0.02
	$N(-15, 10^2)$	10	200	[619, 629]	[629, 633]	0.01	200	[605, 613]	[622, 626]	0.02
		15	200	[629, 640]	[631, 635]	0.01	200	[621, 633]	[627, 632]	0.00
		20	200	[644, 655]	[646, 650]	0.01	200	[638, 651]	[643, 648]	0.00
	$N(-15, 20^2)$	10	200	[494, 508]	[504, 510]	0.01	200	[486, 492]	[509, 513]	0.04
		15	200	[578, 592]	[584, 588]	0.00	200	[607, 612]	[612, 621]	0.01
		20	200	[674, 686]	[670, 673]	0.02	200	[686, 691]	[684, 690]	0.00
	$U[-40, 20]$	10	100	[412, 421]	[451, 454]	0.07	100	[510, 516]	[542, 545]	0.05
		15	100	[568, 577]	[603, 605]	0.05	100	[617, 632]	[642, 649]	0.03
		20	100	[765, 772]	[798, 801]	0.04	100	[740, 749]	[800, 803]	0.06
	$U[-40, 40]$	10	100	[263, 275]	[267, 272]	0.00	200	[446, 450]	[446, 448]	0.00
		15	200	[334, 344]	[354, 361]	0.05	200	[515, 518]	[516, 518]	0.00
		20	200	[475, 492]	[504, 510]	0.04	200	[586, 604]	[645, 649]	0.06
Homo1	$N(0, 10^2)$	10	100	[400, 414]	[433, 442]	0.07	200	[484, 494]	[493, 495]	0.01
		15	100	[545, 553]	[569, 573]	0.04	100	[567, 580]	[586, 588]	0.02
		20	200	[700, 710]	[710, 714]	0.05	200	[695, 701]	[698, 702]	0.00
	$N(-15, 10^2)$	10	200	[524, 532]	[530, 535]	0.01	200	[511, 517]	[524, 527]	0.02
		15	200	[528, 537]	[533, 536]	0.00	200	[524, 535]	[532, 535]	0.00
		20	200	[546, 554]	[549, 552]	0.00	200	[543, 553]	[547, 551]	0.00
	$N(-15, 20^2)$	10	200	[389, 408]	[427, 437]	0.02	200	[454, 460]	[479, 482]	0.01
		15	200	[493, 504]	[500, 504]	0.00	200	[[520, 525]	527, 534]	0.02
		20	200	[594, 597]	[593, 602]	0.00	200	[608, 612]	[607, 612]	0.00
	$U[-40, 20]$	10	200	[377, 384]	[383, 385]	0.01	200	[442, 447]	[462, 465]	0.03
		15	200	[514, 521]	[521, 524]	0.01	200	[544, 555]	[551, 553]	0.00
		20	200	[705, 713]	[708, 711]	0.04	200	[705, 713]	[708, 711]	0.00
	$U[-40, 40]$	10	100	[229, 236]	[230, 235]	0.00	200	[373, 381]	[393, 395]	0.04
		15	100	[377, 384]	[383, 385]	0.01	200	[442, 447]	[462, 465]	0.03
		20	100	[483, 495]	[498, 502]	0.02	200	[634, 638]	[645, 647]	0.02
Homo2	$N(0, 10^2)$	10	200	[599, 613]	[611, 617]	0.01	200	[702, 717]	[716, 719]	0.01
		15	200	[780, 797]	[788, 795]	0.00	200	[824, 833]	[812, 820]	0.02
		20	200	[884, 901]	[890, 897]	0.00	200	[884, 892]	[898, 904]	0.01
	$N(-15, 10^2)$	10	200	[793, 805]	[795, 802]	0.00	200	[771, 780]	[787, 792]	0.02
		15	200	[793, 806]	[798, 804]	0.00	200	[791, 807]	[800, 805]	0.00
		20	200	[793, 806]	[799, 805]	0.00	200	[791, 807]	[800, 805]	0.01
	$N(-15, 20^2)$	10	200	[590, 603]	[636, 644]	0.07	200	[668, 676]	[703, 708]	0.05
		15	200	[708, 724]	[721, 728]	0.01	200	[756, 765]	[745, 753]	0.01
		20	200	[779, 794]	[783, 789]	0.00	200	[798, 805]	[804, 812]	0.01
	$U[-40, 20]$	10	200	[447, 459]	[459, 464]	0.01	200	[486, 492]	[509, 513]	0.04
		15	200	[717, 728]	[730, 735]	0.01	200	[721, 744]	[731, 742]	0.00
		20	200	[907, 920]	[913, 918]	0.00	200	[872, 884]	[920, 926]	0.04
	$U[-40, 40]$	10	200	[321, 327]	[326, 332]	0.02	200	[558, 561]	[558, 562]	0.00
		15	100	[392, 406]	[411, 420]	0.04	200	[615, 620]	[614, 618]	0.01
		20	100	[486, 507]	[516, 524]	0.04	200	[670, 675]	[683, 686]	0.01

3.7.5 Computation Time of the TSM-AO and TSM-PI

Table 3.8 present range of the average solution time for the TSM-AO and TSM-PI (reported as average solution time with N=5 scenarios–200 scenarios) across 20 random instances for each combination of patient mix, stochastic arrival distribution, and grace period. First, we note that the TSM-AO enjoys an outstanding computational behavior independent of patient mix (equivalently, distributions of service durations), distribution of arrival time deviations, length of the grace period, and the cost structure. Solution time of the TSM-AO ranges from 0.01 seconds (with $N =$ 5 scenarios) to 66 seconds (with $N = 200$ scenarios), demonstrating its implementability (i.e., can be easily translated into standard optimization software packages, not requiring customized algorithmic development or tuning).

Second, the computational behavior of the TSM-PI depends on patient mix, distribution of arrival time deviations, and length of the grace period. For example, the TSM-PI takes longer times to solve the heterogeneous (Hetero) instances than homogeneous (Homo) instances. The TSM-PI solution time (in seconds) ranges from 0.01 (with $N = 5$ scenarios) to 7200 (with $N = 200$ scenarios) and from 0.01 (with $N = 5$ scenarios) to 3700 (with $N = 200$ scenarios) for the Hetero and Homo instances, respectively. Moreover, the TSM-PI solution time increases as the variability of arrival time deviation increases and as the grace period shortens. For example, for Hetero1 and a grace period of 10 minutes, increasing the variability from $u \sim N(-15, 10^2)$ to $u \sim N(-15, 20^2)$ widens the range of the TSM-PI solution time from 0.2–13 to 15–5312 seconds, respectively. In contrast, increasing the grace period from 10 to 20 minutes under $u \sim N(-15, 20^2)$ narrowed the range of the TSM-PI solution time from 15–5312 to 0.2–17 seconds.

We attribute the longer solution time of the TSM-PI for the instances with heterogeneous patient types, higher variability in arrival time deviations, and shorter grace periods to the following. First, heterogeneous instances require larger number of binary decision variables representing the initial appointment sequencing, which leads to larger MIP models. Second, under higher variability in arrival time deviation, a larger portion need to be rescheduled (i.e., arrive within interval R or DE). This results in larger search space of the rescheduling decisions in the second-stage formulation of the TSM-PI.

Finally, we note that, without strengthening the TSM-PI (see Section 3.4.2), we were unable to solve any of the 60 TSM-PI instances or even smaller instances with 3 patients and 10 scenarios. As we solve the TSM-PI only to obtain a statistical lower bound for SOASP and evaluate the (sub-)optimality of the appointment order policy, we did not attempt to further improve the solution efficacy of TSM-PI.

Table 3.8: The range of the average solution time for the TSM-AO and TSM-PI (reported as average solution time with N=5 scenarios–200 scenarios) across 20 random instances for each combination of patient mix, stochastic arrival distribution, and grace period. Large solution time are marked in bold.

Inst	**G**	**N(0,10^2)**		**N(-15,10^2)**		**N(-15,20^2)**		**U[-40,20]**		**U[-40,40]**	
		TSM-AO	**TSM-PI**	**TSM-AO**	**TSM-PI**	**TSM-AO**	**TSM-PI**	**TSM-AO**	**TSM-PI**	**TSM-AO**	**TSM-PI**
Hetero1	**10**	0.39–23	66–**7214**	0.02–8	0.2–13	0.2–13	15–**5312**	0.02–2	1–**3615**	0.02–0.7	12–*
	15	0.23–39	2–**1885**	0.31–7	0.02–9	10–53	3–1852	0.02–3	1–219	0.02–1	9–*
	20	0.2–22	0.5–175	0.2–7	0.2–8	0.2–16	0.2–17	0.02–2	0.02–0.7	0.01–1	5–*
Hetero2	**10**	1–41	1–**2647**	0.2–0.3	0.2–22	0.3–66	0.3–**628**	0.02–2	2–**3623**	0.03– 2	7–*
	15	0.3–35	1–58	0.3–0.4	0.26–13	0.3–38	0.2–55	0.02–2	2–341	0.04–1	4–*
	20	0.4–1	0.3–15	0.2–0.4	0.3–10	0.3–4	0.2–3	0.01–1	0.02–1	0.02–1	1–*
Homo1	**10**	0.03–0.7	0.2–**1612**	0.01–0.5	0.4–4	0.02–1	0.2–**2188**	0.01–1.2	11-**3249**	0.02–3	5–**3601**
	15	0.01–1	0.1–46	0.02–0.5	0.02–0.4	0.01–0.5	0.08–30	0.02–1	1–73	0.01–7	15–**3666**
	20	0.01–0.3	7–10	0.02–0.4	0.01–0.4	0.01–0.6	0.06–0.6	0.01–0.3	0.05–0.4	0.03–2	12–795
Homo2	**10**	0.01–0.7	6–**1950**	0.02–0.3	0.02–0.5	0.01–4	0.2–214	0.02–0.3	10–**3269**	1–3	14–**3700**
	15	0.01–0.7	0.07–55	0.01–0.6	0.02–0.7	0.01–0.5	0.14–55	0.02–1	2–175	1–5	10–**3500**
	20	0.02–0.3	0.06–0.7	0.01–0.3	0.01–0.7	0.01–0.3	0.1–0.6	0.01–1	0.5–2	1–4	10–**3409**

[*] terminated with 1% relative MIP Gap (`relMIPGap` $:= \frac{UB-LB}{UB} \times 100\%$, where UB is the best upper bound and LB is the linear programming (LP) relaxation-based lower bound obtained at termination after 2 hours) at the time limit of 2 hours. We note that we allowed such instances to run for several days, however, the relative MIP gap remained at 1% for some instances and 5% for all others. For these instances, we use the optimal value of the LP relaxation of TSM-PI at termination instead of v^{PI} in computing the ARG% in Section 3.5.3.

3.7.6 Best Selection of Rescheduling Period R^* in NS policy

Table 3.9: Values of the rescheduling period R^* associated with the tight ARG values under the NS policy in Figures 3.2–3.3.

			Hetero1/Hetero2					
G	**N(0,10^2)**		**N(-15,20^2)**		**U[-40,20]**		**U[-40,40]**	
	Cost1	Cost2	Cost1	Cost2	Cost1	Cost1	Cost1	Cost2
10	10	5	5	5	10	5	10	5
15	N/A	N/A	N/A	N/A	N/A	3	5	5
20	N/A	N/A	N/A	N/A	N/A	N/A	5	3
			Homo1					
G	**N(0,10^2)**		**N(-15,20^2)**		**U[-40,20]**		**U[-40,40]**	
	Cost1	Cost2	Cost1	Cost2	Cost1	Cost1	Cost1	Cost2
10	10	5	10	5	5	3	15	5
15	N/A	N/A	N/A	N/A	N/A	N/A	10	5
20	N/A	N/A	N/A	N/A	N/A	N/A	5	3
			Homo2					
G	**N(0,10^2)**		**N(-15,20^2)**		**U[-40,20]**		**U[-40,40]**	
	Cost1	Cost2	Cost1	Cost2	Cost1	Cost1	Cost1	Cost2
10	10	5	10	5	5	5	10	5
15	N/A	N/A	N/A	N/A	N/A	N/A	5	5
20	N/A	N/A	N/A	N/A	N/A	N/A	5	3

3.7.7 Average Number of Declined Appointments Under the NS and AO Policies

Table 3.10: Average number (per day) of declined appointments under the NS and AO policies

		$N(0,10^2)$			
Inst	**G**	Cost1		Cost2	
		NS	AO	NS	AO
Hetero1/2	10	0	2	0	2
Homo1/2	10	0	2	1	2
		$N(-15,20^2)$			
Inst	**G**	Cost1		Cost2	
		NS	AO	NS	AO
Hetero1/2	10	0	2	0	2
Homo1/2	10	0	1	0	2
		$U[-40,20]$			
Inst	**G**	Cost1		Cost2	
Hetero1/2	10	0	2	0	2
Homo1/2	10	1	2	1	2
		$U[-40,40]$			
Inst	**G**	Cost1		Cost2	
		NS	AO	NS	AO
Hetero1/2	10	3	5	3	5
	15	2	4	2	4
	20	1	2	1	2
Homo1	10	2	5	3	5
	15	2	4	2	4
	20	1	3	2	3
Homo2	10	3	5	3	5
	15	2	4	2	4
	20	1	3	1	3

3.7.8 Improvement in Scheduling Metric

Table 3.11: The average and maximum improvement of S^{u} over S^{0} (reported as average–maximum) in scheduling metrics under various patient mixes, stochastic arrival distributions, and Cost 1 (i.e.,$c^w = c^g = c^o$). Results are based on 20 random instances for each parameter setting.

Inst	**Distribution**	**Obj(%)**	**WaitT(%)**	**OverT(%)**	**IdleT(%)**
Hetero1	**N(0,10^2)**	18–31%	20–30%	41–65%	8–16%
	N(-15,10^2)	10–24%	17–38%	2–24%	22–33%
	N(-15,20^2)	17–26%	13–21%	35–54%	18–24%
	U[-40,20]	7–20%	8–23%	16–46%	8–16%
	U[-40,40]	26–28%	6–9%	198–217%	38–39%
Hetero2	**N(0,10^2)**	11–20%	12–25%	29–43%	4–8%
	N(-15,10^2)	7–15%	9–20%	5–24%	11–18%
	N(-15,20^2)	9–19%	11–23%	13–32%	9–15%
	U[-40,20]	9–18%	10–21%	27–47%	3–9%
	U[-40,40]	38–46%	37–51%	2–18%	53–60%
Homo1	**N(0,10^2)**	15–17%	7–11%	39–42%	21–23%
	N(-15,10^2)	7–30%	17–49%	-4–23%	8–11%
	N(-15,20^2)	7–15%	7–13%	29–35%	18–22%
	U[-40,20]	7–18%	10–25%	6–19%	11–17%
	U[-40,40]	26–35%	9–21%	40–67%	51–59%
Homo2	**N(0,10^2)**	13–16%	12–21%	30–41%	11–17%
	N(-15,10^2)	9–27%	24–54%	6–20%	-2–9%
	N(-15,20^2)	10–23%	20–37%	1–20%	5–12%
	U[-40,20]	8–18%	13–26%	8–24%	3–8%
	U[-40,40]	28–32%	1–10%	88–96%	44–51%

Table 3.12: The average and maximum improvement of S^{u} over S^{0} (reported as average–maximum) in scheduling metrics under various patient mixes and stochastic arrival distributions, and Cost 2 (i.e.,$c^w = 1,\ c^g = 5,\ c^o = 7.5$). Results are based on 20 random instances for each parameter setting.

Inst	Distribution	Obj(%)	WaitT(%)	OverT(%)	IdleT(%)
Hetero1	**N(0,10^2)**	7–18%	13–21%	18–20%	7–7%
	N(-15,10^2)	1–14%	2–21%	4–26%	0–15%
	N(-15,20^2)	6–20%	1–11%	14–43%	19–31%
	U[-40,20]	7–18%	-8–0%	21–49%	33–48%
	U[-40,40]	40–46%	-21– -15%	113–162	124–132
Hetero2	**N(0,10^2)**	8–21%	-8–4%	15-45%	43-54%
	N(-15,10^2)	6–22%	3–33%	-16–29	-9–7%
	N(-15,20^2)	7–22	-1–11%	13–50%	20–30%
	U[-40,20]	10–20%	-7–7%	22–25%	40–43%
	U[-40,40]	45–52%	-14%–8%	127–187%	104–112%
Homo1	**N(0,10^2)**	3–17%	3–14%	7–21%	5-13%
	N(-15,10^2)	3–15%	6–21%	0–28%	4–11%
	N(-15,20^2)	6–19	0–13%	13–42%	17–27%
	U[-40,20]	7–17%	-2–8%	15–36%	31–39%
	U[-40,40]	29–42%	-24–17%	99–641%	119–142%
Homo2	**N(0,10^2)**	10–21%	-6–2%	20–31%	45–48%
	N(-15,10^2)	7–22%	6–30%	-16–30%	-17–12%
	N(-15,20^2)	8–20%	2–11%	21–45%	17–29%
	U[-40,20]	10–32%	13–32%	12–62%	6–16%
	U[-40,40]	27–40%	-39– -33%	79–197	123–145%

CHAPTER 4

A Distributionally Robust Optimization Approach for Outpatient Colonoscopy Scheduling

4.1 Introduction

In this chapter, we consider an OPC manager who must schedule the start time for a set of colonoscopy procedures (appointments) for a single provider. Colonoscopy duration and patient actual arrival time relative to their scheduled appointment time are random in nature and observed on the day of service, after the appointment decisions are made. The quality of a schedule's performance is a function of patient waiting time, provider idle time, and provider overtime.

This chapter is based on our work with the University of Michigan Medical Procedure Unit (UM-MPU), an OPC that performs a variety of procedures including a large number of colonoscopies. Colonoscopy, in particular, is the mainstay of diagnosis and prevention for colorectal cancer (CRC), a leading cause of cancer death worldwide (*Anderson and Butterly* (2015), American Cancer Society 2019, *Singh et al.* (2016); *Zauber et al.* (2012))

Colonoscopy appointment planning decisions are challenging for several reasons. First, there is significant variability in colonoscopy duration, primarily due to the quality of pre-procedure bowel preparation (prep) that the patient must undergo (*Bechtold et al.* (2016); *Chokshi et al.* (2012); *Froehlich et al.* (2005); *Johnson et al.* (2014); *Lebwohl et al.* (2010); *Rex et al.* (2002, 2006)). Our analysis of the UM-MPU data suggests that colonoscopy durations are "*bimodal*", i.e., depending on the prep quality they can follow two different probability distributions, one for those with adequate prep and the other for those with inadequate prep (see this analysis in Section 4.2). Unfortunately, when scheduling a patient, it is not known at that time, whether the patient will perform an adequate prep or not. Furthermore, there is a wide range of possible probability distributions for modeling the variability in colonoscopy duration with adequate and inadequate prep (see Figure 4.1 in Section 4.2).

Second, colonoscopy is often scheduled with upper endoscopy. The variability in the duration of the combined colonoscopy and upper endoscopy is primarily due to the variability in colonoscopy duration (as a function of uncertain prep quality). Moreover, the duration of a combined procedure is longer than that of a colonoscopy procedure. This requires the OPC managers to make complex sequencing decisions pertaining to the order of colonoscopy and the combined upper endoscopy and colonoscopy.

Third, several clinical studies suggest that time of the day may affect colonoscopy outcomes, possibly as a consequence of provider fatigue as the day progresses (see, e.g., *Almadi et al.* (2015); *Singh et al.* (2016)). As such, the provider often has a preference for earlier start times for those who are at high risk of CRC. Accommodating provider preference and maintaining good operational performance are difficult to trade off. For example, scheduling the combined procedure of a high-risk patient at the start of the day may increase the waiting time of the subsequent appointment.

Finally, there is significant variability in patient actual arrival time relative to their scheduled arrival time (see Figure 4.6 in Appendix 4.7.1).

Ignoring the variability of colonoscopy duration can lead to patient delay, provider idling and/or overtime. By incorporating uncertainty, classical stochastic appointment scheduling models, as well as those that we propose in chapters 2–3, seek to find scheduling decisions that minimize the expected cost of patient waiting, provider idle time, and provider overtime

In this chapter, we exploit the ideas and tools of distributionally robust (DR) optimization to address uncertainty in both procedure duration (as a function of uncertain prep quality) and arrival time deviations. We consider DR outpatient colonoscopy scheduling (DROCS) problem that seeks optimal appointment sequence and schedule to minimize the worst-case expected weighted sum of patient waiting, provider idling, and provider overtime. Here, we take the worst-case over an ambiguity set (a family of distributions) characterized by the known means and supports of prep quality, durations, and arrival time deviations. We derive an equivalent mixed-integer linear programming (MILP) formulation to solve DROCS.

Using the UM-MPU data, we then conduct extensive numerical experiments to draw insights into colonoscopy scheduling. Specifically, we demonstrate that this DR approach can produce schedules that (1) have a good operational performance (in terms of waiting time, idle time, and overtime) under various probability distributions (and extreme scenarios) of the random parameters, and (2) can accommodate provider (and patient) preference on appointment time while maintaining a good operational performance as compared to the stochastic programming approach.

To the best of our knowledge, and according to the recent review of outpatient appointment systems by *Ahmadi-Javid et al.* (2017) and the literature review in Section 4.3, the work in this chapter is the first to address the bimodal ambiguity of colonoscopy durations. We further contribute with a new DR model that incorporates sequencing decisions and considers the ambiguity of two coexisting uncertainties of colonoscopy duration (as a function of uncertain prep quality) and arrival time deviation.

The remainder of this chapter is structured as follows. In the Section 4.2, we present a motivating example for our DR approach. In Section 4.3, we review the relevant literature. In Section 4.4, we formally define DROCS and its MILP reformulation. In Section 4.5, we use historical colonoscopy data to conduct case studies of DROCS and draw managerial insights based on real data. Finally, we conclude and summarise this chapter in Section 4.6.

Table 4.1: The following parameter estimates (in minutes) are based on historical data from the Gastroenterology and Hepatology Endoscopy Practice at University of Michigan Medical Procedures Unit between 2013 and 2017.

Parameter	Mean	Standard Deviation
Arrival time deviation, u	10	25
Colonoscopy duration with adequate prep,d^{A}	21	11
Colonoscopy duration with inadequateprep, d^{I}	25	15
Prep Adequacy rate $\simeq$ 0.86		

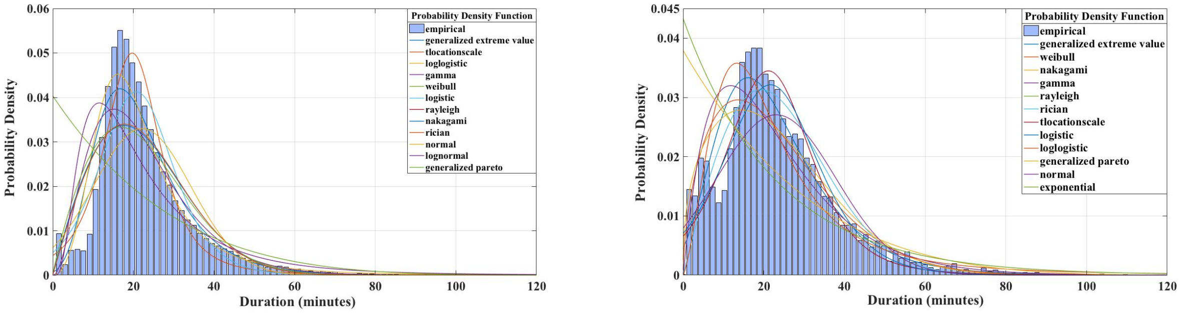
4.2 Motivation

A main challenge in scheduling appointment times for colonoscopy procedures at the UM-MPU is the variability in colonoscopy duration due to adequate versus inadequate bowel prep. To measure this variability and analyze these distributional differences, we analyzed UM-MPU data collected electronically from January 2013 through December 2017, which represents ∼45K colonoscopy appointments. Table 4.1 provides statistics on the mean and standard deviation of colonoscopy durations and the deviations of arrival time from scheduled appointment time.

We analyze the probability distribution of colonoscopy durations with adequate and with inadequate prep using the following statistical tests at a significance level $\alpha = 0.05$.

1. *Hartigan's dip test of unimodality* (*Hartigan et al.*, 1985): for a random vector X having distribution F, this test examines the null hypothesis that F is a unimodal distribution. Correspondingly, the alternative hypothesis is that F is non-unimodal, i.e., at least bimodal. We implemented this test using the R function `dip.test` (*Maechler*, 2016) with X representing the vector of all colonoscopy durations. We obtain a p-value of $2.2\times10^{-16} < \alpha$, rejecting the null hypothesis in favor of the alternative hypothesis, i.e., the distribution of colonoscopy durations is non-unimodal and so is at least bimodal.

2. *Two-sample Kolmogorov-Smirnov (KS) test*: This test determines if two random vectors Y and Z differ probabilistically without making any assumptions about their distributions.



(a) Actual colonoscopy duration with adequate prep (b) Actual colonoscopy duration with inadequate prep

Figure 4.1: The empirical and fitted probability distributions for colonoscopy duration with (a) adequate bowel prep and (b) inadequate bowel prep.

The null hypothesis is that Y and Z have the same distribution; and the alternative hypothesis is that Y and Z have different distributions. We implemented the KS test using the open source Matlab function `kstest2` with Y and Z representing the data vectors of colonoscopy durations with adequate and inadequate prep, respectively. We obtain a *p*-value of $2.80 \times 10^{-33} < \alpha$ and so we reject the null hypothesis in favor of the alternative hypothesis that the distribution of colonoscopy durations with adequate prep is different than with inadequate prep

3. Candidate distributions: we used the Matlab function `allfitdist` (*Sheppard*, 2012) to fit all parametric distributions to the data vectors of colonoscopy durations with adequate and inadequate prep. For a given data vector, `allfitdist` returns the fitted distributions and goodness-of-fit metrics (e.g., Negative of the Log Likelihood (NLogL), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC)). Figure 4.1 and the values of NLogL, AIC, BIC in Appendix 4.7.2 demonstrates that a wide range of distributions can well represent colonoscopy duration with adequate (inadequate) prep.

These statistical results motivate us to model the colonoscopy durations with adequate and inadequate prep with different distributions. In addition, the fact that colonoscopy durations can be well represented by a wide range of distributions motivates us to adopt a DR approach.

4.3 Literature Review

In this section, we focus primarily on the literature on stochastic appointment scheduling. For comprehensive surveys of outpatient appointment scheduling, we refer the reader to *Ahmadi-Javid et al.* (2017), *Cayirli and Veral* (2003), and *Gupta and Denton* (2008). More broadly, *Pinedo* (2016) provides a detailed survey of a wide range of scheduling problems, including their theory, algorithms, and applications. For the clinical literature on the relationship between bowel prep quality and colonoscopy duration, we refer to *Anderson and Butterly* (2015); *Bechtold et al.* (2016); *Chokshi et al.* (2012); *Froehlich et al.* (2005); *Johnson et al.* (2014); *Lebwohl et al.* (2010); *Rex et al.* (2002, 2006).

Within the stochastic appointment scheduling literature, most studies focus on uncertainty pertaining to service duration (in part because it is the primary source of disruption in clinic operations) and few papers study and incorporate the variability in arrival time (*Ahmadi-Javid et al.*, 2017). Furthermore, most studies that consider scheduling multiple patients types often assume that the service duration of each type follows one probability distribution (see, e.g., *Berg et al.* (2014) and the references therein). Therefore, this chapter is one of the few studying outpatient scheduling with bimodal service duration.

For appointment scheduling with stochastic service duration, we refer to *Begen et al.* (2012); *Bosch and Dietz* (2000); *Berg et al.* (2014); *Cayirli and Yang* (2014); *Denton and Gupta* (2003); *Ge et al.* (2013); *Shehadeh et al.* (2019); *Robinson and Chen* (2003); *Mittal et al.* (2014) and the references therein. For appointment scheduling with random arrival time we refer to *Alexopoulos et al.* (2008); *Cayirli and Veral* (2003); *Deceuninck et al.* (2018); *Glowacka et al.* (2017); *Klassen and Yoogalingam* (2014); *Samorani and Ganguly* (2016) and the references therein.

Simulation and stochastic programming (SP) are by far the most common approaches to deal with uncertainty in appointment scheduling problems. Both approaches assume that random parameters follow fully known distributions. Simulation models aim at developing and evaluating different scheduling heuristics (see, e.g., *Ahmadi-Javid et al.* (2017); *Cayirli et al.* (2006, 2008); *Gul et al.* (2011); *Klassen and Yoogalingam* (2014) and the references therein). On the other hand, SP models often aim at finding scheduling decisions that minimize the expected cost of waiting, provider idling, and provider overtime, where the expectation is taken with respect to a probability distribution of random parameters that is assumed to be known (see, e.g., *Shehadeh et al.* (2019); *Robinson and Chen* (2003)).

When the probability distributions of random parameters are not known, an SP approach can lead to poor scheduling decisions. As pointed out by *Esfahani and Kuhn* (2018), if we calibrate an SP model to a given data sample and evaluate its optimal decisions on a different data sample, then the resulting out-of-sample performance is often disappointing. This phenomenon is known as the optimizers' curse, i.e., an attempt to optimize based on imperfect estimates of probability distributions leads to biased decisions (see *Esfahani and Kuhn* (2018) and *Smith and Winkler* (2006) for detailed discussions). Furthemore, as pointed out by *Esfahani and Kuhn* (2018) and *Hanasusanto et al.* (2016), evaluating the objective function of an SP often involves taking multi-dimensional integrals, which is #P-hard. Hence, SP scheduling models can be challenging to solve.

Distributionally robust (DR) optimization is an alternative approach for decision making under uncertainty when the probability distribution governing the uncertain problem data is hard to characterize and so itself subject to uncertainty (*Wiesemann et al.*, 2014). DR scheduling models aim at finding scheduling decisions that minimize the worst-case expected cost of scheduling metrics, where the worst-case is taken over an ambiguity set. The ambiguity set is a family of distributions characterized by some known properties of the unknown probability distributions of random parameters (*Esfahani and Kuhn*, 2018).

In this paper, we adopt a DR model to study colonoscopy scheduling under uncertainty. We consider an ambiguity set based on the probability of having adequate/inadequate bowel prep, as well as the first-moment and support information of colonoscopy durations and arrival time deviations. We refer to *Bertsimas and Popescu* (2005), *Bertsimas et al.* (2010), *Delage and Ye*

(2010), *Scarf* (1958) and references therein for a thorough discussion of DR optimization using moment-based ambiguity sets, and to *Gabrel et al.* (2014) for a thorough review of recent advances in robust and DR optimization.

For the DR appointment scheduling literature, we refer the reader to the pioneering work by *Jiang et al.* (2017), *Kong et al.* (2013), *Kong et al.* (2015), *Mak et al.* (2014), and *Zhang et al.* (2017). *Kong et al.* (2013) and *Mak et al.* (2014) point out that DR models can yield appointment schedules that maintain good performance under various probability distributions and extreme scenarios of random parameters. *Kong et al.* (2013) consider a cross-moment ambiguity set and derive a convex conic programming reformulation of the DR. *Mak et al.* (2014) consider a marginal-moment ambiguity set and derive tractable reformulations based on linear program and second-order conic program.

Jiang et al. (2017) generalize the DR appointment scheduling model of *Mak et al.* (2014) by incorporating heterogeneous no-shows and their distributional ambiguity along with that of service duration. Although this results in a challenging mixed-integer nonlinear reformulation, and *Jiang et al.* (2017) develop integer programming approaches, including valid inequalities, to effectively accelerate the computation of the DR model.

In this chapter, we study a DR model to incorporate the bimodal service duration due to random bowel prep. To the best of our knowledge, this is the first work to address the bimodal ambiguity in colonoscopy appointment scheduling. We further extend this DR model to incorporate sequencing decisions. By reformulating this DR model as a MILP, we provide an implementable tool to obtain insights into outpatient colonoscopy scheduling. Note that, different from *Jiang et al.* (2017), we do not consider random no-shows here because they are not frequently observed at our collaborating OPC[1].

[1] The no-show rate there is very low during 2013–2017. This is reasonable because colonoscopy patients need to fast and go through bowel prep before their appointments. It is hence physically costly to not show up for their appointments.

4.4 DROCS Formulation and Analysis

4.4.1 Random Parameters and Assumptions

We consider sequencing a set of P procedures (patients) for a single provider and determining the associated scheduled time for each procedure. The procedure duration of a patient is random and depends on the prep quality, which is also random when scheduling appointments. In addition, a patient may not arrive exactly at the scheduled time and the arrival time deviation (i.e., the difference between the scheduled and the actual arrival times) is random. The joint probability distribution of all random parameters (prep quality, procedure durations, and arrival time deviations) is assumed ambiguous. We make the following assumptions on DROCS:

A1. The provider is always available at the scheduled start time of the first procedure, and immediately after finishing each procedure.

A2. The provider serves patients in the order of their scheduled appointments, regardless of their actual arrival times (a standard assumption in the offline appointment scheduling, see, e.g., *Deceuninck et al.* (2018) and the references therein).

A3. All the scheduled patients show up to their appointments (no-show is not frequently observed at the collaborating OPC; see the footnote on page 90).

A4. Rescheduling during the day and accommodating walk-ins or emergencies are not permitted. This a standard assumption in the offline stochastic appointment scheduling (see, e.g., *Berg et al.* (2014); *Denton et al.* (2007)) that mimics the practice of the collaborating OPC.

4.4.2 Modeling Procedure Duration as a Function of Bowel Prep

We model the prep quality by a 0-1 Bernoulli random variable q_p such that $q_p = 1$ if the bowel prep of patient p is adequate, and $q_p = 0$ otherwise for all $p = 1, \dots, P$. Accordingly, the procedure duration of each patient p equals $q_p d_p^{\text{A}} + (1 - q_p) d_p^{\text{I}}$, where d_p^{A} and d_p^{I} represent the random procedure duration with adequate and inadequate prep, respectively, for all $p = 1, \dots, P$.

4.4.3 Modeling Scheduling Metrics under Uncertainty

Let binary decision variable $x_{p,i}$ represent the assignment of patient p to appointment i (equivalently, position i in the sequence), for all $i, p = 1, \dots, P$. Let t_i represent the scheduled time of appointment i for all $i = 1, \dots, P$. The feasible region $\mathcal{X}$ of variables x is defined in (4.1) such that each patient is assigned to one appointment and each appointment is assigned one patient. The feasible region $\mathcal{T}$ of variables t is defined in (4.2) such that all appointments are scheduled within the provider service hours $[0, L]$, and the first appointment is scheduled at the start of the day, i.e., $t_1 = 0$.

$$\mathcal{X} = \left\{ x : \begin{array}{c} \sum_{p=1}^{P} x_{p,i} = 1, \forall i = 1, \dots, P, \\ \sum_{i=1}^{P} x_{p,i} = 1, \ \forall p = 1, \dots, P, \\ x_{p,i} \in \{0, 1\}, \forall i, p = 1, \dots, P \end{array} \right\} \tag{4.1}$$

$$\mathcal{T} = \Big\{ t : \ t_1 = 0, \ 0 \leq t_i \leq \mathcal{L}, \ \ \forall i = 1, \dots, P, \ t_i \geq t_{i-1}, \ \forall i = 2, \dots, P \Big\} \tag{4.2}$$

Due to random procedure durations and arrival times, one or multiple of the following scenarios may happen: (i) delay on the start time of an appointment due to late completion of the previous appointments, (ii) idleness of the provider due to early finish of an appointment or tardiness of the next appointment, and (iii) provider overtime beyond his/her $\mathcal{L}$ to finish serving all appointments.

Table 4.2: Notation.

Indices	
p	index of patient, or procedure, $p = 1, ..., P$
i	index of positions in the sequence, or appointments, $i = 1, ..., P$
Parameters	
c_i^{w}	unit waiting time cost of appointment i
c_i^{g}	unit provider idle time cost between appointments $i - 1$ and i
c^{o}	unit provider overtime cost
$\mathcal{L}$	scheduled service hours of the provider
u_p	arrival time deviation of patient p
q_p	probability of adequate bowel prep of patient p
d_p^{A}	procedure duration with adequate prep of patient p
d_p^{I}	procedure duration with inadequate prep of patient p
First-stage decision variables	
t_i	scheduled start time of appointment i
$x_{p,i}$	binary assignment variable indicating whether procedure p is assigned to appointment i
Second-stage decision variables	
w_i	waiting time of appointment i, for all $i = 1, \ldots, P$
g_1	provider idle time before the first appointment
g_i	provider idle time between appointmentss $i - 1$ and i, for all $i = 2, \ldots, P$
w_{P+1}	provider overtime

For all $i = 1, \ldots, P$, let the continuous decision variables a_i and w_i represent the actual arrival time and the waiting time of appointment i, respectively. Let the continuous decision variable w_{P+1} represents provider overtime. For all $i = 2, \ldots, P$, let the continuous decision variable g_i represents provider idle time between the completion of appointment $i - 1$ and the arrival of appointment i, and let decision variable g_1 represents provider idle time before the first appointment. For all $p = 1, \ldots, P$, let the continuous parameter u_p represents the arrival time deviation (i.e., the difference between the scheduled and the actual arrival times) of patient p. Table 4.2 summarizes these notation. Given a feasible schedule ($x \in \mathcal{X}$, $t \in \mathcal{T}$) and a joint realization of uncertain parameters $(q, d^{\text{A}}, d^{\text{I}}, u)$, we can compute patients arrival times $a = [a_1, \ldots, a_P]^\top$, patient waiting times $w = [w_1, \ldots, w_P]^\top$, provider idle time $g = [g_1, \ldots, g_P]^\top$, and provider overtime w_{P+1} using the following recursions.

$$a_i = t_i + \sum_{p=1}^{P} u_p x_{p,i}, \quad \forall i = 1, \dots, P \tag{4.3a}$$

$$w_1 = \max\big\{ - a_1, 0\big\}, \tag{4.3b}$$

$$w_i = \max\{0, [a_{i-1} + w_{i-1} + \sum_{p=1}^{P}(q_p d_p^{\text{A}} + (1-q_p)d_p^{\text{I}})x_{p,i-1}] - a_i\}, \ \forall i = 2, \dots, P \tag{4.3c}$$

$$g_1 = \max\{a_1, 0\}, \tag{4.3d}$$

$$g_i = \max\{0, a_i - [a_{i-1} + w_{i-1} + \sum_{p=1}^{P}(q_p d_p^{\text{A}} + (1-q_p)d_p^{\text{I}})x_{p,i-1}]\}, \ \forall i = 2, \dots, P \tag{4.3e}$$

$$w_{P+1} = \max\{0, [a_P + w_P + \sum_{p=1}^{P}(q_p d_p^{\text{A}} + (1-q_p)d_p^{\text{I}})x_{p,P}] - \mathcal{L}\}. \tag{4.3f}$$

where $a_i + w_i$ is the actual start time of appointment i, $\sum_{p=1}^{P}(q_p d_p^{\text{A}} + (1-q_p)d_p^{\text{I}})x_{p,i}$ is procedure duration of appointment i, and $a_i + w_i + \sum_{p=1}^{P}(q_p d_p^{\text{A}} + (1-q_p)d_p^{\text{I}})x_{p,i}$ is the completion time of appointment i. Let the non-negative parameters c_i^{w}, c_i^{g}, and c^{o} represent the unit penality costs of waiting, idling, and overtime, for all $i = 1, \dots, P$. We formulate the following linear program (LP) to compute the total cost of waiting, idling, and overtime for a given feasible schedule (x, t) and realization of the random parameters $\xi := [q, d^A, d^I, u]^\top$

$$Q(x, t, \xi) := \min_{w,a,g} \sum_{i=1}^{P}(c_i^{\text{w}} w_i + c_i^{\text{g}} g_i) + c^{\text{o}} w_{P+1} \tag{4.4a}$$

$$\text{s.t. } a_i = t_i + u_i, \qquad \forall i = 1, \dots, P \tag{4.4b}$$

$$w_1 - g_1 + a_1 = 0 \tag{4.4c}$$

$$w_i - g_i = a_{i-1} + w_{i-1} + q_{i-1}d_{i-1}^{\text{A}} + (1-q_{i-1})d_{i-1}^{\text{I}} - a_i, \ \forall i = 2, \dots, P \tag{4.4d}$$

$$w_{P+1} - g_{P+1} = a_P + w_P + q_P d_P^{\text{A}} + (1-q_P)d_P^{\text{I}} - \mathcal{L}, \tag{4.4e}$$

$$(w_i, g_i) \geq 0, \ \forall i = 1, \dots, P+1 \tag{4.4f}$$

The objective function (4.4a) minimizes a linear cost function of waiting, idling, and overtime. Constraints (4.4b) compute the actual arrival time as the scheduled time plus the arrival time deviation. If the patient scheduled in appointment i is punctual (early), then $u_i = (<)\ 0$ and so $a_i = (<)\ t_i$. If the patient scheduled in appointment i is late, then $u_i > 0$ and so $a_i > t_i$. Constraint (4.4c) yields either provider idle time before the arrival of the first appointment or the waiting time of the first appointment. Constraints (4.4d) yield either the waiting time of appointment i or the provider idle time between appointment $i-1$ and i based on the arrival time of appointment i and the completion time of appointment $i-1$, i.e., $a_{i-1} + w_{i-1} + q_{i-1}d^{\text{A}}_{i-1} + (1-q_{i-1})d^{\text{I}}_{i-1}$. Constraint (4.4e) yields either the overtime or the schedule earliness. Finally, constraint (4.4f) specify feasible ranges of the decision variables.

4.4.4 Ambiguity Set and DR Model

Classical two-stage scheduling models (and those that we propose in chapters 2–3) seek to find a schedule (x, t) that minimizes the expectation of the random cost $Q(x, t, \xi)$ subject to uncertainty $(q, d^{\text{A}}, d^{\text{I}}, u)$ with a known joint probability distribution denoted as $\mathbb{P}$. In our problem, we assume that $\mathbb{P}$ is not perfectly known. We, however, know the support (i.e., upper and lower bound) and the mean values of the random parameters.

Note that clinical guidelines limit the range of colonoscopy durations. Several empirical studies, including our analysis of the UM-MPU data, suggest that the arrival time deviations are bounded (see, e.g., *Deceuninck et al.* (2018) and Figure 4.6 in Appendix 4.7.1). Therefore, we assume that the clinic manager can estimate from historical data the lower and upper bounds of these random parameters. Mathematically, we consider support $S = S^{\text{q}} \times S^{\text{A}} \times S^{\text{I}} \times S^{u}$, where S^{q}, S^{A}, S^{I}, and S^{u} are respectively the supports of random parameters q, d^{A}, d^{I}, and u defined as follows:

$$\mathcal{S}^{\text{q}} := \{0, 1\}^{P},$$

$$\begin{aligned}
\mathcal{S}^{\mathrm{u}} &:= \{u : u_p^{\mathrm{L}} \leq u \leq u_p^{\mathrm{U}},\ \forall p = 1, \dots, P,\ u_{P+1} = 0\}, \\
\mathcal{S}^{\mathrm{I}} &:= \{d^{\mathrm{I}} \geq 0 : d_p^{\mathrm{IL}} \leq d_p^{\mathrm{I}} \leq d_p^{\mathrm{IU}},\ \forall p = 1, \dots, P,\ d_{P+1}^{\mathrm{I}} = 0\}, \\
\mathcal{S}^{\mathrm{A}} &:= \{d^{\mathrm{A}} \geq 0 : d_p^{\mathrm{AL}} \leq d_p^{\mathrm{A}} \leq d_p^{\mathrm{AU}},\ \forall p = 1, \dots, P,\ d_{P+1}^{\mathrm{A}} = 0\}.
\end{aligned}$$

In addition, we let μ^{q}, μ^{A}, μ^{I}, and μ^{u} represent the mean values of q, d^{A}, d^{I}, and u, respectively. We denote $\mu := \mathbb{E}_{\mathbb{P}}[\xi] = [\mu^{\mathrm{q}}, \mu^{\mathrm{A}}, \mu^{\mathrm{\ I}}, \mu^{\mathrm{u}}]^\top$ for notational brevity. Then, we consider the following mean-support ambiguity set $\mathcal{F}(S, \mu)$:

$$\mathcal{F}(S, \mu) := \left\{ \mathbb{P} \in \mathcal{P}(S) : \begin{array}{l} \int_S d\mathbb{P} = 1 \\ \mathbb{E}_{\mathbb{P}}[\xi] = \mu \end{array} \right\} \tag{4.5}$$

where $\mathcal{P}(S)$ in $\mathcal{F}(S, \mu)$ represents the set of probability distributions supported on S and each distribution matches the mean values of q, d^{A}, d^{I}, and u. Note that in $\mathcal{F}(S, \mu)$ we do not consider higher moments (e.g., covariance, correlation, etc.) of (q, d^{A}, d^{I}, u) for several reasons. First, several studies have shown that service durations and arrival times are independent (*Deceuninck et al.* (2018)). Second, even if the service duration and arrival times are dependent, it is difficult for the clinical manager to accurately estimate the correlation between these uncertain parameters, especially when data is limited. Third, incorporating higher moments can undermine the computational tractability of the DR model, and so its implantability in practice (*Mak et al.*, 2014; *Jiang et al.*, 2017). Using the ambiguity set $\mathcal{F}(S, \mu)$, we formulate the DROCS as the following min-max problem:

$$\text{(DROCS)} \quad \min_{x,t} \ \sup_{\mathbb{P} \in \mathcal{F}(S, \mu)} \ \mathbb{E}_{\mathbb{P}}[Q(x, t, \xi)] \tag{4.6}$$

which searches for a scheduling decision (x, t) that minimizes the worst-case expected cost of waiting, idling, and overtime over a family of distributions characterized by the ambiguity set $\mathcal{F}(S, \mu)$.

4.4.5 Reformulations

In this section, we reformulate the DROCS model in (4.6) into one that is solvable via a commercial solver. In Section 4.4.5.1, we derive an exact LP reformulation with a fixed sequence. Then, in Section 4.4.5.2, we incorporate the sequencing decisions and derive an equivalent MILP formulation for the DROCS model.

4.4.5.1 LP Reformulation with a Fixed Sequence

In this section, we analyze the DR model in (4.6) and derive an exact LP reformulation with a fixed sequence. We first consider the inner maximization problem $\sup\limits_{\mathbb{P} \in \mathcal{F}(S,\mu)} \mathbb{E}_{\mathbb{P}}[Q(x,t,\xi)]$ for a fixed schedule ($x \in \mathcal{X}, t \in \mathcal{T}$), where $\mathbb{P}$ is the decision variable, i.e., we are choosing the distribution that maximizes the expected value of the random cost $Q(x,t,\xi)$. For a fixed (x, t), we can formulate this inner maximization problem as the following linear functional optimization problem.

$$\max \; \mathbb{E}_{\mathbb{P}}[Q(x,t,\xi)] \tag{4.7a}$$

$$\text{s.t.} \quad \mathbb{E}_{\mathbb{P}}[\xi] = \mu, \tag{4.7b}$$

$$\mathbb{E}_{\mathbb{P}}[\mathbb{1}_S(\xi)] = 1 \tag{4.7c}$$

where $\mathbb{1}_S(\xi)$ represents the indicator function of set S such that $\mathbb{1}_S(\xi) = 1$ if $\xi \in S$ and $\mathbb{1}_S(\xi) = 0$ if $\xi \notin S$. Our LP reformulation of (4.7) is inspired by the work of *Mak et al.* (2014) and relies directly on the special structure of the random cost function $Q(x,t,\xi)$, defined in (4.4), as we discuss next. Taking the dual of $Q(x,t,\xi)$ leads to the following proposition (see Appendix 4.7.3 for the proof).

Proposition 4.4.1. *For fixed* $(x \in \mathcal{X}, t \in \mathcal{T})$*, it holds that*

$$Q(x,t,\xi) = \max_{y \in Y} \sum_{i=1}^{P} (t_i + u_i + q_i d_i^A + (1-q_i) d_i^I) y_{i+1} - \sum_{i=1}^{P} (t_i + u_i) y_i - \mathcal{L} y_{P+1}. \tag{4.8}$$

where y*:=*$[y_1, \dots, y_{P+2}]^\top$ *and*

$$Y = \{y_{P+2} = 0,\ c_i^w + y_{i+1} \geq y_i \geq -c_i^g, \textit{ for } i = 1, ..., P+1\},\ c_{P+1}^w = c^o, \textit{ and } c_{P+1}^g = 0. \tag{4.9}$$

In view of equation (4.8), problem (4.7) is equivalent to

$$\begin{aligned} &\max\ \mathbb{E}_{\mathbb{P}}\Big[\max_{y \in Y} \sum_{i=1}^{P} (t_i + u_i + q_i d_i^{\mathrm{A}} + (1-q_i) d_i^{\mathrm{I}}) y_{i+1} - \sum_{i=1}^{P} (t_i + u_i) y_i - \mathcal{L} y_{P+1} \Big] && \text{(4.10a)} \\ &\text{s.t. } (4.7\text{b}) - (4.7\text{c}). && \text{(4.10b)} \end{aligned}$$

As shown in the proof of Proposition 4.4.2 in Appendix 4.7.4, the stochastic optimization problem (4.10) is equivalent to the deterministic problem (4.11).

Proposition 4.4.2. *For any* $(x \in \mathcal{X}, t \in \mathcal{T})$*, problem* (4.10) *is equivalent to*

$$\min_{\rho,\alpha,\lambda,\gamma} \left\{ \sum_{i=1}^{P} \mu_i^A \rho_i + \mu_i^I \alpha_i + \mu_i^u \lambda_i + \mu_i^q \gamma_i + \max_{y \in Y} h(x,t,y,\rho,\alpha,\lambda,\gamma) \right\} \tag{4.11}$$

where

$$\begin{aligned} h(x,t,y,\rho,\alpha,\lambda,\gamma) := \max_{(q,d^A,d^I,u) \in S} \Big\{ &\sum_{i=1}^{P} (t_i + u_i + q_i d_i^A + (1-q_i) d_i^I) y_{i+1} - \sum_{i=1}^{P} \big((t_i + u_i) y_i \\ &- \mathcal{L} y_{P+1}\big) - \sum_{i=1}^{P} (d_i^A \rho_i + d_i^I \alpha_i + u_i \lambda_i + q_i \gamma_i) \Big\} \end{aligned} \tag{4.12}$$

It is straightforward to see that function $h(x,t,y,\rho,\alpha,\lambda,\gamma)$ is convex in variables y. Hence,

$\max_{y \in Y} h(x, t, y, \rho, \alpha, \lambda, \gamma)$ is a convex maximization problem. It follows from the fundamental convex analysis (see, e.g., *Boyd and Vandenberghe* (2004)) that there exists an optimal solution y^* to the inner maximization problem $\max_{y \in Y} \; h(x, t, y, \rho, \alpha, \lambda, \gamma)$ in (4.11) at one of the extreme point of the polyhedron Y defined in (4.9). This motivates us to follow a similar approach to *Mak et al.* (2014) in deriving an equivalent LP reformulation of (4.11) (or equivalently, formulation (4.7)) as follows. First, using the properties of the extreme point of Y, we derive an equivalent LP reformulation of $\max_{y \in Y} \; h(x, t, y, \rho, \alpha, \lambda, \gamma)$. Then, we can reformulate the min-max problem (4.11) as a convex minimization problem. We formally prove this in the following proposition.

Proposition 4.4.3. *The optimal objective value of* $\max_{y \in Y} \; h(x, t, y, \rho, \alpha, \lambda, \gamma)$ *in* (4.11) *is equal to*

$$\min_{\beta} \sum_{i=1}^{P+2} \beta_i \tag{4.13a}$$

$$\begin{aligned}
\textit{s.t.} \sum_{i=1}^{j} \beta_i \geq{}& (-t_1 - u_1^L)\pi_{1,j} + \sum_{i=2}^{\min\{j,P\}} \Big(- t_i - u_i^L + t_{i-1} + u_{i-1}^L + \min\{d_{i-1}^{AL}, d_{i-1}^{IL}\}\Big)\pi_{i,j} \\
&+ \sum_{i=2}^{\min\{j,P+1\}} \Big(\max\{d_{i-1}^{AU}, d_{i-1}^{IU}\} - \min\{d_{i-1}^{AL}, d_{i-1}^{IL}\}\Big)(\pi_{i,j})^+ + \sum_{i=1}^{\min\{j,P\}} K_i' \\
&+ \sum_{i=P+1}^{\min\{j,P+1\}} \Big(t_P + u_P^L - \mathcal{L} + \min\{d_P^{AL}, d_P^{IL}\}\Big)\pi_{P+1,j}, \quad \forall j = 1, \ldots, P+2
\end{aligned} \tag{4.13b}$$

$$\begin{aligned}
\sum_{i=k}^{j} \beta_i \geq{}& (u_{k-1}^U - u_{k-1}^L)(\pi_{k,j} + c_{k-1}^g)^+ + \sum_{i=P+1}^{\min\{j,P+1\}} \Big(t_P + u_P^L - \mathcal{L} + \min\{d_P^{AL}, d_P^{IL}\}\Big)\pi_{P+1,j} \\
&+ \sum_{i=\min\{k,P+1\}}^{\min\{j,P\}} \Big(- t_i - u_i^L + t_{i-1} + u_{i-1}^L + \min\{d_{i-1}^{AL}, d_{i-1}^{IL}\}\Big)\pi_{i,j} + \sum_{i=\min\{k,P+1\}}^{\min\{j,P\}} K_i' \\
&+ \sum_{i=\min\{k,P+1\}}^{\min\{j,P+1\}} \Big(\max\{d_{i-1}^{AU}, d_{i-1}^{IU}\} - \min\{d_{i-1}^{AL}, d_{i-1}^{IL}\}\Big)(\pi_{i,j})^+, \forall k = 2, \ldots, P+1, \\
&\forall j = k, \ldots, P+2
\end{aligned} \tag{4.13c}$$

$$\beta_{P+2} \geq 0, z_i \geq 0, \; z_i \geq \rho_i, \; v_i \geq 0, \; v_i \geq \alpha_i, \; r_i \geq 0, \; r_i \geq \lambda_i, \; e_i \geq 0, \; e_i \geq -\gamma_i, \; \forall i \leq P \tag{4.13d}$$

where for all $i = 1, \dots, P$, $K'_i = -\Big(d_i^{AU}\rho_i + (d_i^{AL} - d_i^{AU})z_i\Big) - \Big(d_i^{IU}\alpha_i + (d_i^{IL} - d_i^{IU})v_i\Big) - \Big(u_i^{U}\lambda_i + (u_i^{L} - u_i^{U})r_i\Big) + e_i$.

As we show in the proof of Proposition 4.4.3 in Appendix 4.7.5, the definition of $\pi_{i,j}$ is motivated by the characterization of extreme points of Y. Specifically, for any extreme point $y_i \in Y$ and for any $1 \leq i \leq j \leq P+2$, $y_i = \pi_{i,j} = -c^g_j + \sum_{\ell=i}^{j-1} c^{\mathrm{w}}_\ell$ and $y_{P+2} = \pi_{P+2,P+2} = 0$. Substituting $\max\limits_{y \in Y} h(x,t,y,\rho,\alpha,\lambda,\gamma) = \min\{\sum_{i=1}^{P+2} \beta_i : \text{(4.13b)} - \text{(4.13d)}\}$ in formulation (4.11), we obtain the following LP reformulation for the DROCS with fixed sequence $x \in \mathcal{X}$.

$$\min_{t \in \mathcal{T}, \rho, \alpha, \lambda, \gamma, \beta} \quad \sum_{i=1}^{P} \mu_i^{\mathrm{A}}\rho_i + \mu_i^{\mathrm{I}}\alpha_i + \mu_i^{\mathrm{u}}\lambda_i + \mu_i^{\mathrm{q}}\gamma_i + \sum_{i=1}^{P+2} \beta_i \tag{4.14a}$$

$$\text{s.t.} \qquad \text{(4.13b)} - \text{(4.13d)}. \tag{4.14b}$$

4.4.5.2 MILP Reformulation with Sequencing

Recall that our DR model in (4.6) incorporates both the sequencing and scheduling decision. The DR-LP formulation in (4.14), however, assumes a fixed sequence. To determine the sequencing and scheduling decisions jointly, we multiply the mean, lower bound, and upper bound of each random parameter in (4.14) with the associated sequencing decisions. This lead to an equivalent MINLP reformulation of the DROCS problem. Due to its large size, we present this formulation in Appendix 4.7.6 and highlight that this formulation has a nonlinear objective function $\sum\limits_{i=1}^{P}\sum\limits_{p=1}^{P} \mu_p^{\mathrm{A}} x_{p,i}\rho_i + \mu_p^{\mathrm{I}} x_{p,i}\alpha_i + \mu_p^{\mathrm{u}} x_{p,i}\lambda_i + \mu_p^{\mathrm{q}} x_{p,i}\gamma_i + \sum\limits_{i=1}^{P+2} \beta_i$ and also nonlinear terms $K''_i = -\Big(\sum\limits_{p=1}^{P} \big(d_p^{\mathrm{AU}} x_{p,i}\rho_i + (d_p^{\mathrm{AL}} - d_p^{\mathrm{AU}}) x_{p,i} z_i\big)\Big) - \Big(\sum\limits_{p=1}^{P} \big(d_p^{\mathrm{IU}} x_{p,i}\alpha_i + (d_p^{\mathrm{IL}} - d_p^{\mathrm{IU}}) x_{p,i} v_i\big)\Big) - \Big(\sum\limits_{p=1}^{P} \big(u_p^{\mathrm{U}} x_{p,i}\lambda_i + (u_p^{\mathrm{L}} - u_p^{\mathrm{U}}) x_{p,i} r_i\big)\Big) + e_i$ in constraints (4.25b)–(4.25c).

To linearize this MINLP formulation, we define $\eta_{p,i} = x_{p,i}\rho_i$, $\tau_{p,i} = x_{p,i}\alpha_{p,i}$, $\Lambda_{p,i} = x_{p,i}\lambda_i$, $\Gamma_{p,i} = x_{p,i}\gamma_i$, $\zeta_{p,i} = x_{p,i}z_i$, $\nu_{p,i} = x_{p,i}v_i$, and $\varphi_{p,i} = x_{p,i}r_i$. We also introduce McCormick inequalities (4.26a)–(4.26g) for variables $\zeta_{p,i}$, $\nu_{p,i}$, $\varphi_{p,i}$, $\eta_{p,i}$, $\tau_{p,i}$, $\Lambda_{p,i}$, and $\Gamma_{p,i}$ respectively. We formally introduce the resulting MILP reformulation of DROCS in Proposition 4.4.4. Note that the McCormick inequalities often rely on big-M coefficients that take large values and can undermine the computational efficiency. In Appendix 4.7.7, we derive tight bounds of these big-M coefficients to strengthen the MILP formulation.

Proposition 4.4.4. *The following MILP is equivalent to the DROCS problem in* (4.6)

$$\begin{aligned}
(\textit{DR-bimodal}) \quad & \min \ \sum_{i=1}^{P}\sum_{p=1}^{P} \mu_p^A \eta_{p,i} + \mu_p^I \tau_{p,i} + \mu_p^u \nu_{p,i} + \mu_p^q \Gamma_{p,i} + \sum_{i=1}^{P+2} \beta_i && \text{(4.15a)} \\
& \textit{s.t.} \ \ (x \in \mathcal{X}, t \in \mathcal{T}) && \text{(4.15b)} \\
& \qquad \text{(4.25b)} - \text{(4.25d)}, \ \text{(4.26d)} - \text{(4.26g)} && \text{(4.15c)}
\end{aligned}$$

4.5 Computational Results

In this section, we compare our DR colonoscopy scheduling approach with the stochastic programming approaches and draw several insights into colonoscopy scheduling. Specifically, we construct several DROCS instances based on the current UM-MPU practice and compare the optimal schedules of DR-bimodal in (4.15) with those yielded by: (1) SMILP-bimodal, a SP approach that considers random prep quality (see Appendix 4.7.8 for the formulation), (2) DR-plain, a DR model that ignores the random prep quality (see Appendix 4.7.9 for the formulation), and (3) SMILP-plain, a SP approach that ignores the random prep quality (see Appendix 4.7.8 for the formulation). In the plain models, we ignore prep adequacy and assume that colonoscopy durations follow a single probability distribution.

We summarize our computational study as follow. We first follow a distributional belief to generate N independent and identically distributed (i.i.d.) samples of each random parameter. Second, we compute the support and mean information from the generated samples and use them to obtain the (in-sample) optimal solutions and optimal objective values to the DR models. Third, we solve the SMILPs using the generated sample and compare (1) optimal sequencing and scheduling patterns yielded by the DR and the SMILP models, and (2) the in-sample and out-of-sample performance of the optimal schedules of the DR and SMILP. Finally, we use the DR model to study the value of incorporating bowel prep prediction in colonoscopy scheduling.

In Section 4.5.1, we describe the DROCS instances that we use in our experiments and discuss the experiment setups. In Section 4.5.2, we compare the optimal DR and SMILP scheduling patterns (i.e., the structure of the optimal time assigned for colonoscopy appointments). In Section 4.5.3, we analyze the optimal DR and SMILP sequencing decisions for the colonoscopy procedure and combined upper endoscopy and colonoscopy procedure. In Section 4.5.4, we evaluate the cost of revising the optimal sequence of the DR and SMILP to accommodate provider preference. In Section 4.5.5, we compare the out-of-sample performance of the optimal schedules of the DR and SMILP models. Finally, in Section 4.5.6, we study the value of incorporating prep adequacy prediction in colonoscopy scheduling.

4.5.1 Description of Experiments

Our computational study is based on the Gastroenterology and Hepatology Endoscopy Practice at the UM-MPU. Specifically, we consider two DROCS instances each with $P = 10$ patients, reflecting the typical daily schedule of the majority of UM-MPU providers. The first instance is homogenous consisting of 10 colonoscopy procedures. The second instance is heterogeneous consisting of 6 colonoscopy procedures (C) and 4 combined upper endoscopy and colonoscopy (UC) procedures. Table 4.3 provides statistics on the mean and standard deviation of procedure durations and arrival time deviations of the UM-MPU appointments during the period from January 2, 2013, to December 15, 2017.

We use our data and follow the same procedure as in prior appointment scheduling studies (see, e.g., *Jiang et al.* (2017); *Denton and Gupta* (2003); *Mak et al.* (2014)) to generate random param-

Table 4.3: The following parameter estimates (in minutes) are based on historical data from the Gastroenterology and Hepatology Endoscopy Practice at University of Michigan Medical Procedures Unit between 2013 and 2017.

Parameter	**Mean**	**Standard Deviation**
Arrival time deviation, u	10	10
Colonoscopy duration with adequate prep, d^{A}	21	11
Colonoscopy duration with inadequate prep, d^{I}	25	15
Upper endoscopy and colonoscopy duration with Adequate prep, d^{UCA}	26	13
Upper endoscopy and colonoscopy duration with Inadequate prep, d^{UCI}	29	15
Prep adequacy rate $\simeq$ 0.86		

eters for each DROCS instance as follows. We set the mean μ^{A} (μ^{I}) and the standard deviation σ^{A} (σ^{I}) of colonoscopy duration with adequate (inadquate) prep to their empirical values, i.e., $\mu_p^{\text{A}} = 21$, $\mu_p^{\text{I}} = 25$, $\sigma^{\text{A}} = 12$, and $\sigma^{\text{I}} = 15$, for all $p = 1, \ldots, P$. Similarly, we set the mean and standard deviation of UC duration with adequate and inadequate prep to their empirical values, i.e., $\mu_p^{\text{UCA}} = 26$, $\mu_p^{\text{UCI}} = 29$, $\sigma^{\text{UCA}} = 13$, and $\sigma^{\text{UCI}} = 15$, for all $p = 1, \ldots, P$. We also set the mean μ^{u} and standard deviation σ^{u} of arrival time deviations to their empirical values, i.e., $\mu_p^{\text{u}} = 10$ and $\sigma_p = 10$, for all $p = 1, \ldots, P$. Finally, we set $\mu_p^{\text{q}} = 0.86$ for all $p = 1, \ldots, P$, which reflect the prep adequacy rate at UM-MPU in the period of 2013–2017.

To approximate the lower (d^{AL}, d^{IL}, u^{L}) and upper (d^{AU}, d^{IU}, u^{U}) bounds of $(d^{\text{A}}, d^{\text{I}}, u)$, we respectively use the 20%-quantile and 80%-quantile values of the N in-sample data. We generate the in-sample colonoscopy durations with adequate prep, d^{A}, and inadequate prep, d^{I}, by following lognormal (LogN) distributions. Specifically, we sample N =1000 realizations (d_1^{An}, $\ldots, d_P^{\text{An}}), \ldots, (d_1^{\text{AN}}, \ldots, d_P^{\text{AN}})$ from LogN(μ_p^A, σ_p^A) and N =1000 realizations $(d_1^{\text{I1}}, \ldots, d_P^{\text{I1}}), \ldots, (d_1^{\text{IN}}$, $\ldots, d_P^{\text{IN}})$ from LogN(μ_p^I, σ_p^I) using the generated means (μ_p^A, μ_p^I) and standard deviations (σ_p^A, σ_p^I) of (d_p^{A}, d_p^{I}) for each $p = 1, \ldots, P$. The lognormal distribution is one of the candidate distributions for colonoscopy durations with adequate/inadequate prep (see Appendix 4.7.2) and a typical distribution to model service duration in appointment scheduling literature (see, e.g., Cayirli et al.

(2006); *Gul et al.* (2011) and references therein). We generate $(q_1^n, \dots, q_P^n)$, $n = 1, \dots, N$, from Bernoulli distribution with mean $\mu^{\mathrm{q}} = 0.86$ (prep adequacy rate at UM-MPU in the period of 2013–2017).

We similarly generate the in-sample UC procedure durations data by following LogN distributions with (μ_p^{UCA}, μ_p^{UCI}) and (σ_p^{UCA}, σ_p^{UCI}). For DR-plain, we first combine all historical C (UC) durations data and obtain the mean of the resulting data vector. We then follow the same procedure described above to generate the in-sample data and estimate the support information for each random parameter. To generate the in-sample arrival time deviations, we sample $N = 1000$ relizations $u_1^n, \dots, u_P^n$, $n = 1, \dots, N$, by following a normal distribution with the generated mean and standard deviations of u_p. The normal distribution is a common distribution to model arrival time deviation in the appointment scheduling literature (see, e.g., *Deceuninck et al.* (2018); *Klassen and Yoogalingam* (2014)).

We consider two different cost structures for the objective function: (1) Cost1: $c^w = c^g = c^o$; and (2) Cost2: $c^w = 1,\ c^g = 5,\ c^o = 7.5$. For the first cost, each of the three objectives is equally important (a classical assumption in the domain of appointment scheduling, see, e.g., *Berg et al.* (2014); *Deceuninck et al.* (2018); *Shehadeh et al.* (2019)). The second cost structure fixes the $c^{\mathrm{o}}/c^{\mathrm{g}}$ ratio to 1.5 as in *Deceuninck et al.* (2018), based on OPC practice (see, e.g., *Cayirli et al.* (2006); *Deceuninck et al.* (2018) for detailed discussions).

Finally, we set the provider service hours based on the current practice of UM-MPU. Providers there typically allocate 30 minutes for each colonoscopy procedure and 45 minutes for the combined UC procedure. Accordingly, we set $\mathcal{L} = 30 \times 10 = 300$ minutes for the homogenous instance and $\mathcal{L} = 30 \times 6 + 45 \times 4 = 360$ minutes for the heterogeneous instance.

For each cost structure, we optimize the SMILP model with the generated N scenarios of each random parameter and the DR model with the generated mean and support of each random parameter. We implemented the DR and SMILP models using the AMPL2016 Programming language calling CPLEX V12.6.2 as a solver with default settings (we didn't observe any consistent benefits in any parameter tuning). We ran all experiments on an HP workstation running Windows Server 2012 with two Intel E5-2620-v4 processor, each with 8-Cores (16 total), 2.10GHz CPUs, and 128 GB shared RAM.

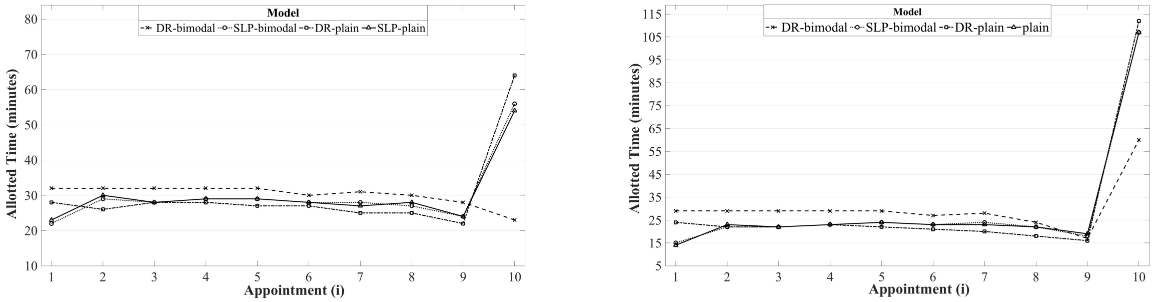
4.5.2 Analysis of the Optimal Scheduling Patterns

In this section, we compare the optimal scheduling patterns (i.e., the structure of the optimal time allowances between appointments) of the DR and SMILP models. We focus on the homogonous instance (i.e., 10 colonoscopy appointments) for which the sequence of appointments is fixed as all procedures are of the same type. In this case, the SMILP models reduce to stochastic linear programs (SLP) and the DR models reduce to linear programs.

Figure 4.2 and Figure 4.3 present the optimal schedules of 10 colonoscopy appointments, produced by the DR and SLP models for the punctual and random arrival cases, respectively. The point (x, y)=(i, allotted time$_i$) of every schedule in each subfigure corresponds to the optimal time assigned for each appointment $i = 1, \dots, 10$ (equivalently, the optimal time interval, $t_{i+1} - t_i$, assigned between the scheduled arrival of appointments i and $i + 1$).

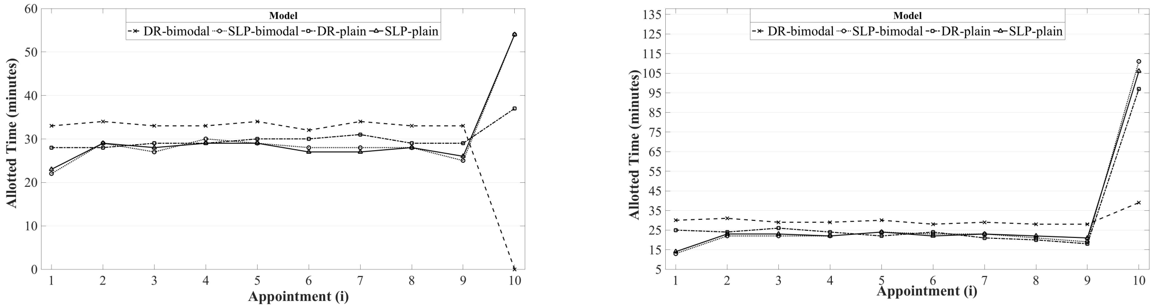
We first observe the following about the optimal DR and SLP scheduling patterns for the punctual arrival case: both SLPs assign less time for the first appointment than the subsequent appointments, equally distribute the time between appointments 2–9, and schedule a longer time for the last appointment than all other appointments. The SLP-bimodal assigns a slightly longer time for each appointment than the SLP-plain. Under Cost2, in which provider time is more important than patient time, both SLPs assign shorter time per appointment and longer time for the last appointment than under Cost1.

As compared to the SLPs, the DR models always assign more time for the first appointment. Under Cost1, the optimal time allowances of the DR-bimodal form a "dome shape," i.e., time



(a) Cost1: $c^w = c^g = c^o$ (b) Cost2: $c^w = 1, c^g = 5, c^o = 7.5$

Figure 4.2: Optimal appointment schedules of the DR and the SLP models with punctual arrivals.



(a) Cost1: $c^w = c^g = c^o$ (b) Cost2: $c^w = 1, c^g = 5, c^o = 7.5$

Figure 4.3: Optimal appointment schedules of the DR and the SLP models with random arrivals.

allowances between appointments first increase then decrease (see Figure 4.2a). Except for the last appointment, the DR-bimodal assigns longer time for each appointment than the DR-plain and the two SLP models under both cost structures. Intuitively, by incorporating the bimodal ambiguity in colonoscopy duration as a function of bowel prep, the DR-bimodal intends to mitigate the waiting time that may accumulate due to long procedure durations with adequate/inadequate prep.

The observations made from the punctual arrival case remain valid in the random arrival case. This indicates that the random arrivals have limited impacts on the optimal schedule patterns for the UM-MPU.

Table 4.4: Optimal Sequencing Patterns with Punctual Arrivals.

		Cost1: $c^{\mathrm{w}} = c^{\mathrm{g}} = c^{\mathrm{o}}$									
		Appointment									
Model		1	2	3	4	5	6	7	8	9	10
DR-bimodal	Type	C	C	C	C	C	UC	C	UC	UC	UC
	Time	29	30	30	31	30	34	30	36	35	77
SMILP-bimodal	Type	C	C	C	C	C	C	UC	UC	UC	UC
	Time	23	27	27	29	27	28	32	33	31	102
DR-plain	Type	C	C	C	C	C	C	UC	UC	UC	UC
	Time	27	27	28	28	27	27	30	30	29	109
SMILP-plain	Type	C	C	C	C	C	C	UC	UC	UC	UC
	Time	21	26	24	25	25	23	29	29	29	129
		Cost2: $c^{\mathrm{w}} = 1, c^{\mathrm{g}} = 5, c^{\mathrm{o}} = 7.5$									
		Appointment									
Model		1	2	3	4	5	6	7	8	9	10
DR-bimodal	Type	C	C	C	UC	UC	C	UC	C	C	UC
	Time	29	27	25	28	28	23	26	19	16	140
SMILP-bimodal	Type	C	C	C	C	C	C	UC	UC	UC	UC
	Time	15	21	22	23	23	23	27	26	23	158
DR-plain	Type	C	C	C	C	C	C	UC	UC	UC	UC
	Time	23	24	22	23	22	21	23	22	22	160
SMILP-plain	Type	C	C	C	C	C	C	UC	UC	UC	UC
	Time	15	21	21	21	22	21	24	25	23	167

4.5.3 Analysis of the Optimal Sequencing Patterns

In this section, we compare the optimal sequencing patterns produced by the DR and SMILP models for the heterogeneous instances. Table 4.4 and Table 4.5 present the optimal sequencing pattern yielded by the DRO and the SMILP models with punctual and random arrivals, respectively.

The DR-plain and the two SMILPs schedule the C procedures before the UC procedures, i.e., sequence procedures by increasing order of mean and variability of procedure duration. In contrast, the optimal sequence of the DR-bimodal always starts with a block of 3-5 consecutive C procedures followed by a block of 1-2 UC procedures, a block of 1-2 C procedures, ending with a block of

Table 4.5: Optimal Sequencing Patterns with Random Arrivals.

		Cost1: $c^{\text{w}} = c^{\text{g}} = c^{\text{o}}$									
		Appointment									
Model		1	2	3	4	5	6	7	8	9	10
DR-bimodal	Type	C	C	C	C	UC	UC	C	C	UC	UC
	Time	34	33	36	34	38	37	33	33	36	46
SMILP-bimodal	Type	C	C	C	C	C	C	UC	UC	UC	UC
	Time	23	29	28	30	29	29	33	33	30	96
DR-plain	Type	C	C	C	C	C	C	UC	UC	UC	UC
	Time	28	30	30	30	29	28	34	35	38	79
SMILP-plain	Type	C	C	C	C	C	C	UC	UC	UC	UC
	Time	23	29	28	30	29	26	33	34	33	95
		Cost2: $c^{\text{w}} = 1, c^{\text{g}} = 5, c^{\text{o}} = 7.5$									
		Appointment									
Model		1	2	3	4	5	6	7	8	9	10
DR-bimodal	Type	C	C	C	C	UC	UC	C	C	UC	UC
	Time	29	29	30	29	30	30	23	21	22	118
SMILP-bimodal	Type	C	C	C	C	C	C	UC	UC	UC	UC
	Time	14	22	22	22	24	23	27	26	23	157
DR-plain	Type	C	C	C	C	C	C	UC	UC	UC	UC
	Time	24	24	24	23	22	19	26	25	21	150
SMILP-plain	Type	C	C	C	C	C	C	UC	UC	UC	UC
	Time	14	23	22	22	24	22	27	26	26	154

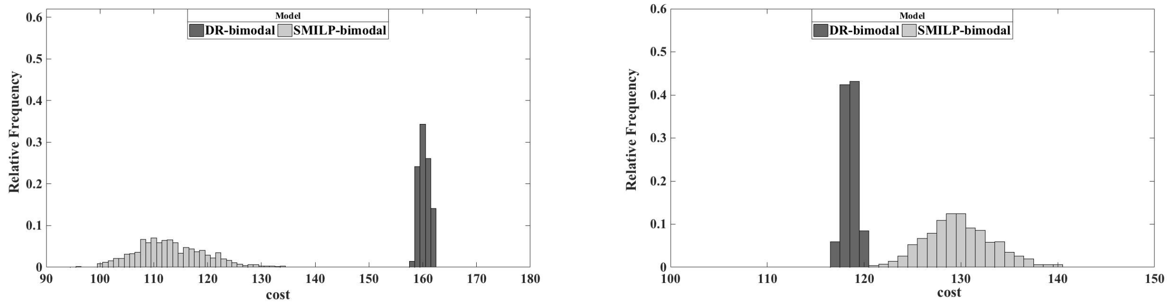
1-3 UC procedures. Moreover, all models assign longer time for the UC procedure than the C procedure (especially, under Cost1), and the DR-bimodal assigns longer time for each procedure type.

Intuitively, by scheduling a slightly longer time per procedure, starting with a block of C procedures and assigning a block of one or two C procedures between the UC procedures (which have larger mean and variability) the DR-bimodal attempts to mitigate the waiting time that may accumulate due to long UC procedure durations.

4.5.4 Cost of Revising the Optimal Sequence

Note that the appointment sequences in Tables 4.4–4.5 are only optimal with respect to the objective of minimizing the scheduling metrics. This objective function does not include the clinical perspective of colonoscopy outcome. Several clinical studies suggest that time of the day may affect colonoscopy outcomes, possibly as a consequence of provider fatigue (see, e.g., *Almadi et al.* (2015); *Singh et al.* (2016) and the references therein). *Almadi et al.* (2015), for example, show that workload before performing a colonoscopy is inversely associated with the detection rates of the precancerous CRC polyps. As such, the provider often has a preference for earlier start times for those who are at high risk of CRC, which may deviate from the optimal sequence of both DR and SMILP model. In this section, we evaluate the potential cost of revising the DR and SMILP optimal sequence when clinical or other perspectives are taken into account.

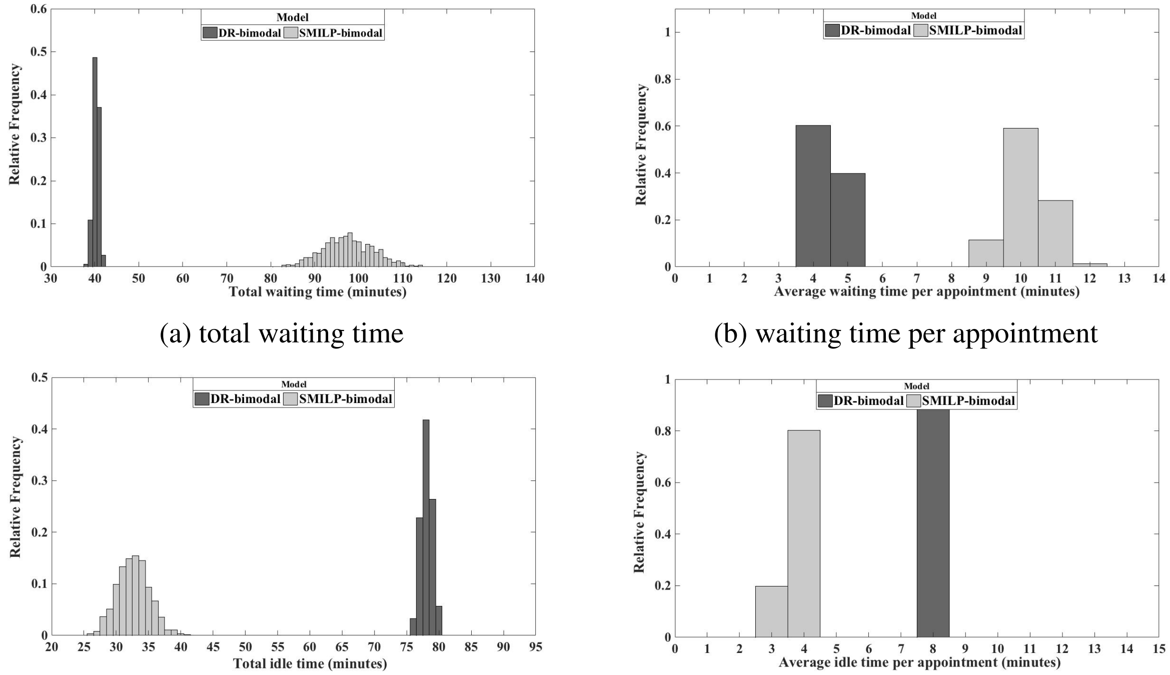
We evaluate the performance of the optimal DR and SMILP schedules for the heterogeneous instances under various sequences as follows. First, we generate 1000 random (and feasible) sequences $\mathbf{X}_1, \dots, \mathbf{X}_{1000}$ of (6C, 4UC), where for all $s = 1, \dots, 1000$, $\mathbf{X}_s \in \mathcal{X}$ defined in (4.1), i.e., each $\mathbf{X}_s$ represents one of the possible feasible assignment of procedures to appointments (see Figures 4.7–4.8 in Appendix 4.7.10). Second, we generate the in-sample data and parameters of the DR model as described in Section 4.5.1. Third, we solve the DR and SMILP models with the generated sequences and the in-sample data. In each solve $s = 1, \dots, 1000$, we fix the sequence in the DR and SMILP models to $\mathbf{X}_s$. Then, we optimize the remaining continuous variables and record the optimal DR and SMILP schedules t_s^{DR} and t_s^{SMILP}, respectively, and the associated in-sample costs (i.e., optimal objective value of the DR and SMILP models). Finally, for all $s = 1, \dots, 1000$, we simulate each t_s^{DR} (t_s^{SMILP}) using a new sample of N'_s =10,000 i.i.d. scenarios of procedure durations and arrival time deviations sampled from the UM-MPU data of each procedure type. That is, we compute $w_i^{n'}$, $g_i^{n'}$, $w_{P+1}^{n'}$ in each scenario $n' = 1, \dots, N'_s$, and the out-of-sample cost $v_s^{\text{DR}}(v_s^{\text{SMILP}}) = (1/N') \sum_{n'=1}^{N'} \sum_{i=1}^{P} w_i^{n'} + g_i^{n'} + w_{P+1}^{n'}$.



(a) Optimal (in-sample) costs

(b) Out-of-sample simulation costs

Figure 4.4: Relative frequency histograms of the optimal (in-sample) and out-of-sample simulation costs.



(a) total waiting time

(b) waiting time per appointment

(c) total provider idle time

(d) provider idle time per appointment

Figure 4.5: Relative frequency histograms of the out-of-sample waiting time and idle time with punctual arrivals.

For presentation brevity, we present and discuss results for the punctual arrival case; our findings for the random arrival case are similar and presented in Appendix 4.7.11. Also, we only present the results for the DR-bimodal and SMILP-bimodal models as they are more relevant to our study. Figure 4.4 present the optimal in-sample costs and out-of-sample simulation costs.

Clearly, the DR-bimodal has smaller fluctuation in the in-sample costs across different sequences than the SMILP-bimodal. In addition, the DR-bimodal has significantly smaller variations in the out-of-sample costs (and smaller values) than the SMILP-bimodal. Moreover, as shown in Figures 4.5, the DR-bimodal has significantly smaller variations in the out-of-sample waiting time and idle time costs (the overtime costs were similar for both models). The DR-bimodal also results in significantly less out-of-sample waiting time and slightly more idle time than the SMILP-bimodal.

We also evaluate the cost of revising (deviating from) the optimal sequence as follows. First, we obtain the minimum out-of-sample cost of the DR and SMILP v^{minDR} and v^{minSP}, respectively, across the 1000 simulation runs. Then, for $s = 1, \dots, 1000$, we evaluate the relative cost gap $\text{CostGap}_s^{\text{DR}}$ ($\text{CostGap}_s^{\text{SMILP}}$) as $\frac{v_s^{\text{DR}}(v_s^{\text{SMILP}}) - v^{\text{minDR}}(v^{\text{minSMILP}})}{v^{\text{minDR}}(v^{\text{minSMILP}})} \times 100\%$. The averages and standard deviations (avg±stdv) of $\text{CostGap}_s^{\text{DR}}$ for the punctual and random arrival cases are $2\% \pm 2$ and $1\% \pm 3$ with a maximum of 3% and 3%, respectively. In contrast, the avg±stdv of $\text{CostGap}_s^{\text{SMILP}}$ for the punctual and random arrival cases are $10 \pm 4\%$ and $10 \pm 5\%$ with a maximum of 17% and 18%, respectively.

These results suggest that OPC can perform, operationally, very well with a random appointment sequence (e.g., in order to accommodate the provider or patient preferences) using the DR schedules. This is a desirable property considering that (1) computation times of finding the "optimal sequence" via SMILP increase quite rapidly as the number of procedures and scenarios of random parameters increases (see Appendix 4.7.12), and (2) in reality, we often need to accommodate the provider preferences and so the implemented appointment sequence may be very different from those obtained from DR-bimodal and SMILP-bimodal models.

4.5.5 Analysis for the Out-of-Sample Performance

In this section, we compare the out-of-sample simulation performance of the optimal schedules of the DR and SMILP models under "*perfect information*" (known distributions) and "*misspecified distribution information*". We generate two sets of N' =10,000 i.i.d. out-of-sample data

$(q_1^n, d_1^{An}, d_1^{In}, u_1^n), \ldots, (q_P^n,\ d_P^{An}, d_P^{In}, u_P^n)$, for $n = 1, ..., N'$ of the random vector $(q, d^{\text{A}}, d^{\text{I}}, u)$ as follows.

1. *Perfect Information*: we use the same distributions and parameter settings as the ones for generating the N in-sample to sample the N' scenarios (i.e., lognormal and normal distributions for procedure durations and arrival time deviations, respectively).
2. *Misspecified Distribution*: we keep the same mean values $(\mu_p^{\text{q}}, \mu_p^{\text{A}}, \mu_p^{\text{I}}, \mu_p^{\text{u}})$ of random parameters $(q_p, d_p^{\text{A}},\ d_p^{\text{I}}, u_p)$ for each patient $p = 1, \ldots, P$. With the mean values kept the same, we vary the distribution of $d_p^{\text{A}}, d_p^{\text{I}}$, and u_i to generate the N' scenarios. This is to simulate the out-of-sample performance of appointment schedules when the in-sample data is biased.

For the case of misspecified distribution, we follow positively correlated truncated normal and weibull distributions with supports $[0, d^{\text{AU}}]$ and $[0, d^{\text{IU}}]$ to generate realizations $(d_1^{\text{An}}, d_1^{In}), \ldots, (d_P^{\text{An}}, d_P^{In}$). We also sample with replacement $(d_1^1, \ldots, d_P^1)$, …, $(d_1^{N'}, \ldots, d_P^{N'})$ from the UM-MPU data and use d_p^n as procedure duration for $p = 1, \ldots, P$ and $n = 1, \ldots, N'$ during the simulation. Furthermore, we follow a positively correlated Bernoulli distribution to generate realizations $q_1^n, \ldots, q_P^n$ and a uniform distribution with support $[u^{\text{L}}, u^{\text{U}}]$ to generate realizations $u_1^n, \ldots, u_P^n$, for $n = 1, \ldots, N'$. We designed the parameters of the normal, weibull, Bernoulli, and uniform distributions to obtain positive data correlations and meanwhile keep the first two moments of the N' out-of-sample realizations the same as the ones of the N in-sample realizations (see, e.g., *Jiang et al.* (2017) and the references therein).

We measure the out-of-sample performance of each optimal schedule (x, t) yielded by DR and SMILP models as follow. First, we fix (x, t) in the SMILP model (see Appendix 4.7.8 for the formulation). Then, we use the out-of-sample parameters (q_1^n, d_1^{An}, d_1^{In}, u_1^n),...,(q_P^n, d_P^{An}, d_P^{In}, u_P^n) to compute w_i^n, g_i^n, and w_{P+1}^n as the waiting time (WT), idle time (IT), and overtime (OT), respectively, in each scenario $n = 1, \ldots, N'$. Tables 4.10–4.11 in Appendix 4.7.13 present means and quantiles of WT (per appointment), IT (per appointment), and OT, yielded by the optimal schedule of the DR and SMLP under perfect distributional information for the homogenous and heterogeneous instances, respectively.

Clearly, the performances of the optimal schedules of the DR-plain and SMILP models are very close on average and at all quantiles. The optimal schedule of the DR-bimodal has shorter waiting time on average and at all quantiles than that of the DR-plain and the SMILP models (which posess perfect distributional information). On the other hand, for the homogenous case, the optimal schedule of the DR-bimodal yields longer overtime under Cost1, and slightly longer overtime under Cost2. Finally, the idle time of the DR and SMILP schedules are comparable. These results imply that when the distributional information is accurate, the DR-bimodal model yields near-optimal appointment schedule.

Tables 4.12-4.15 present the means and quantiles of WT, IT, and OT, yielded by the optimal schedule of the DR and SMILP models under misspecified distributional information. From these results, we observe that the DR-bimodal yields much shorter waiting time per appointment than the DR-plain and the SMILPs schedules when the probability distributions of random parameters are misspecified. The waiting time reductions are significant under both cost structures and are reflected in all quantiles of the random WT. On the other hand, the DR and SMILP schedules yield similar idle time and overtime on average and at the 50%–95% quantiles of random IT and OT.

These observations show how the optimal schedules of the SMILP can become sub-optimal when the probability distributions of random parameters are misspecified, while the DR-bimodal model can produce schedules that more robust (i.e., maintaining good performance under different probability distributions of random parameters). Note that the much shorter waiting time (per appointment) provided by the DR-bimodal is a desirable property for colonoscopy appointments. From the UM-MPU's perspective, minimizing the waiting time of colonoscopy patients is highly desirable given the discomfort that patients experience due to bowel prep (e.g., vomiting, nausea, diarrhea, etc.) and fasting (especially for diabetic patients).

Finally, to quantify the benefits of employing the DR-bimodal schedule in practice, we compare its out-of-sample total cost with that of the current MPU schedule under the MPU data sample. The relative cost gaps=[(Total Cost (DR-bimodal)- Total Cost (Current))/Total cost (DR-bimodal)]$\times 100\%$ for the homogenous instance are -20% and -18% with punctual and random arrivals, respectively. The relative cost gaps for the heterogeneous instance are -106% and -56% with punctual and random arrivals, respectively. These results imply that implementing our DR-bimodal schedules can significantly enhance the clinic operational performance.

4.5.6 Value of Bowel Prep Prediction

In this section, we evaluate the value of perfect prediction of bowel prep adequacy. That is, the value of scheduling appointments based on perfect information about the prep adequacy of all patients.

For illustration purposes, we construct and solve the following three additional homogenous DROCS instances, each with 10 colonoscopies and different probability of adequate bowel prep. Then, we evaluate the cost of scheduling patients according to optimal DR-bimodal schedule for the homogenous base instance instead of scheduling them according to the optimal schedule for each instance.

- Base instance (instb): 10 patients with $\mu^{\mathrm{q}}_p = 0.86$ for all $p = 1, \ldots, 10$ (i.e., assuming homogeneity of patients and ignoring any prediction about prep adequacy).
- Instance 1 (inst1): 5 patients with adequate prep and 5 with inadequate prep, i.e., $q_p \equiv 1$ for $p = 1, \ldots, 5$ and $q_p \equiv 0$ for $p = 6, \ldots, 10$.
- Instance 2 (inst2): 8 patients with adequate prep and 2 with inadequate prep, i.e., $q_p \equiv 1$ for $p = 1, \ldots, 8$ and $q_p \equiv 0$ for $p = 9, 10$.
- Instance 3 (inst3): 2 patients with adequate prep and 8 with inadequate prep, i.e., $q_p \equiv 1$ for $p = 1, 2$ and $q_p \equiv 0$ for $p = 3, \ldots, 10$.

Table 4.6: The relative cost gaps $G^{\text{(instb-inst1)}}$, $G^{\text{(instb-inst2)}}$, and $G^{\text{(instb-inst3)}}$ for the punctual and random arrival cases

	Punctual		**Random**	
	Cost1	**Cost2**	**Cost1**	**Cost2**
$G^{\text{(instb-inst1)}}$	11%	23%	14%	14%
$G^{\text{(instb-inst2)}}$	13%	29%	16%	21%
$G^{\text{(instb-inst3)}}$	10%	23%	11%	12%

We solve instb, inst 1, inst2, and inst3 using the DR-bimodal model, which respectively yeild the schedules S^{instb}, S^{inst1}, S^{inst2}, and S^{inst3}. Then, we compute the cost of ignoring bowel prep prediction in S^{instb} as follows. First, we sample N' =10,000 scenarios of procedure durations d from the UM-MPU data according to inst1 patient mix. That is, we sample $d_1^n, \ldots, d_5^n$, $n = 1, \ldots, N'$ from the data vector of durations with adequate prep and $d_6^n, \ldots, d_{10}^n$, $n = 1, \ldots, N'$ from the data vector of durations with inadequate prep. Then, we simulate S^{instb} and S^{inst1} with the generated sample and compute corresponding cost $C^{\text{(instb-1)}}$ and $C^{\text{(inst1)}}$. Finally, we compute the relative increase in cost $G^{\text{(instb-inst1)}}$ (cost gap between $C^{\text{(instb-1)}}$ and $C^{\text{(inst1)}}$) as $G^{\text{(instb-inst1)}} = \frac{C^{\text{(instb-1)}} - C^{\text{(inst1)}}}{C^{\text{(inst1)}}} 100\%$. We perform the same simulation steps and and relative cost gap calculations for inst2 and inst3.

Table 4.6 presents $G^{\text{(instb-inst1)}}$ and $G^{\text{(instb-inst2)}}$, and $G^{\text{(instb-inst3)}}$. The positive values of these cost gaps indicate that optimizing (customizing) colonoscopy appointment scheduling with a perfect prediction of patient bowel prep adequacy has the potential of reducing the clinic operational costs. Future work will be directed in developing data-driven prediction models of bowel prep adequacy and incorporating these in optimizing colonoscopy appointment scheduling systems.

4.6 Conclusion and Chapter Summary

In this paper, we consider the outpatient colonoscopy scheduling problem, recognizing the impact of prep adequacy on the variability in colonoscopy duration. Data from the UM-MPU indicates that colonoscopy durations are bimodal, i.e., depending on the prep quality they can follow two different probability distributions, one for those with adequate prep and the other for those with inadequate prep. We define a DROCS problem that seeks optimal appointment sequence and schedule to minimize the worst-case weighted expected sum of patient waiting, provider idling, and provider overtime, where the worst-case is taken over an ambiguity set (a family of distributions) characterized through the known mean and support of the prep quality and durations. By deriving an equivalent MILP of the DROCS, we provide an implementable tool to obtain insights into outpatient colonoscopy scheduling.

Using the UM-MPU data, we conduct extensive numerical experiments to draw insights into colonoscopy scheduling. Specifically, we demonstrate that this DR approach can produce schedules that (1) have a good operational performance (in terms of waiting time, idle time, and overtime) under various probability distributions (and extreme scenarios) of the random parameters, and (2) can accommodate provider (and patient) preference on appointment time while maintaining a good operational performance as compared to the stochastic programming approach. We also show that optimizing (customizing) colonoscopy appointment scheduling with a perfect prediction of patient bowel prep adequacy has the potential of reducing the clinic operational costs.

4.7 Appendix

4.7.1 Variability of Patient Arrival Time

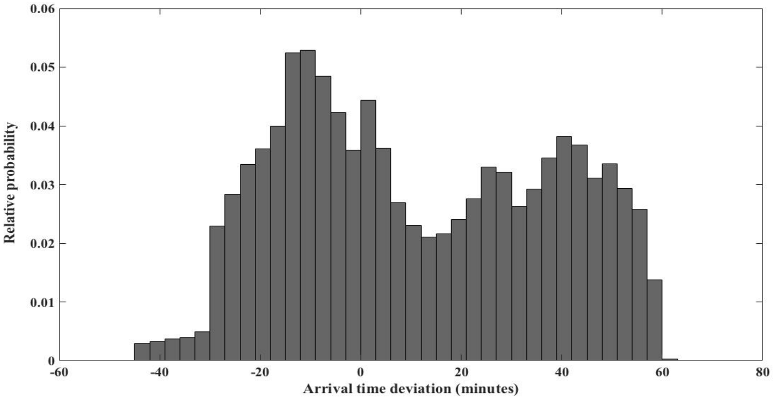


Figure 4.6: Variability of patient arrival time deviation (2013-2017)

4.7.2 Candidate Probability Distributions for Colonoscopy Durations

Table 4.7: Candidate probability distributions for colonoscopy durations with adequate prep and their goodness of fit metrics; Negative of the Log Likelihood (NLogL), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC).

Distribution Name	**NLogL**	**BIC**	**AIC**
generalized extreme value	145161.09	290353.86	290328.17
tlocationscale	146712.79	293457.27	293431.59
loglogistic	146961.88	293944.88	293927.75
gamma	147762.11	295545.35	295528.22
weibull	147885.82	295792.77	295775.65
logistic	147971.33	295963.77	295946.65
rayleigh	147997.13	296004.83	295996.27
nakagami	147991.88	296004.89	295987.77
rician	147997.18	296015.49	295998.36
normal	151111.18	302243.48	302226.35
lognormal	154366.54	308754.21	308737.08
generalized pareto	156309.14	312649.96	312624.27
exponential	158335.06	316680.68	316672.12
extreme value	167355.01	334731.13	334714.01
birnbaumsaunders	180098.90	360218.93	360201.81
inverse gaussian	189926.62	379874.37	379857.25

Table 4.8: Candidate probability distributions for colonoscopy durations with inadequate prep and their goodness of fit metrics; Negative of the Log Likelihood (NLogL), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC).

Distribution Name	**NLogL**	**BIC**	**AIC**
generalized extreme value	22809.89	45645.73	45625.78
weibull	22837.97	45693.24	45679.94
nakagami	22852.68	45722.66	45709.36
gamma	22940.18	45897.65	45884.35
rayleigh	23140.40	46289.45	46282.80
rician	23140.40	46298.10	46284.80
tlocationscale	23157.91	46341.77	46321.82
logistic	23206.54	46430.38	46417.08
loglogistic	23233.96	46485.23	46471.93
generalized pareto	23374.94	46775.82	46755.87
normal	23468.33	46953.96	46940.66
exponential	23639.66	47287.97	47281.32
lognormal	23837.17	47691.64	47678.34
extreme value	25364.14	50745.58	50732.28
birnbaumsaunders	25544.99	51107.28	51093.98
inverse gaussian	26439.78	52896.85	52883.55

4.7.3 Proof of Proposition 4.4.1

Proof. Letting $z = [z_1, \dots, z_P]^\top$ be the dual variables associated with constraints (4.4b), y_1 and $y = [y_2, \dots, y_P]^\top$ be the dual variables associated with constraints (4.4c) and (4.4d), respectively, and y_{P+1} be the dual variable associated with constraint (4.4e), we formulate $Q(x, t, \xi)$ in (4.4) in its dual form as:

$$\max_{z,y} \ \sum_{i=1}^{P}(t_i + u_i)z_i + \sum_{i=1}^{P}(q_i d_i^{\text{A}} + (1 - q_i)d_i^{\text{I}})y_{i+1} - \mathcal{L}y_{P+1} \tag{4.16a}$$

$$\text{s.t.} \quad z_i + y_i - y_{i+1} = 0 \qquad \text{for all } i = 1, \dots, P \tag{4.16b}$$

$$y_i - y_{i+1} \leq c_i^{\text{w}} \qquad \text{for all } i = 1, \dots, P \tag{4.16c}$$

$$- y_i \leq c_i^{\text{g}} \qquad \text{for all } i = 1, \dots, P \tag{4.16d}$$

$$0 \leq y_{P+1} \leq c^{\text{o}} \tag{4.16e}$$

Constraint set (4.16b) is related to primal variables a_i for all $i = 1, \dots, P$, constraint set (4.16c) is related to primal variables w_i for all $i = 1, \dots, P$, constraint (4.16e) is related to primal variables g_{P+1} and w_{P+1}. Note from constraint (4.16b) that the dual variable y entirely determine variable z, i.e., $z_i = y_{i+1} - y_i$ for all $i \in [P]$. Therefore, formulation (4.16) is equivalent to:

$$\max_{y \in Y} \ \sum_{i=1}^{P} (t_i + u_i + q_i d_i^{\text{A}} + (1 - q_i) d_i^{\text{I}}) y_{i+1} - \sum_{i=1}^{P} (t_i + u_i) y_i - \mathcal{L} y_{P+1} \tag{4.17a}$$

$$\text{s.t.} \quad Y := \{(4.16\text{c}) - (4.16\text{e})\} \tag{4.17b}$$

For notational convenience, we define a dummy variable y_{P+2} which always takes the lower-bound value $y_{P+2} = -c^{g}_{P+2} = 0$, $c^{\text{w}}_{P+1} = c^{\text{o}}$, and $c^{\text{g}}_{P+1} = 0$. Accordingly, we rewrite Y as follows:

$$Y = \{y_{P+2} = 0, \ c^{\text{w}}_i + y_{i+1} \geq y_i \geq -c^{\text{g}}_i \text{ for } i = 1, ..., P + 1\} \tag{4.18}$$

□

4.7.4 Proof of Proposition 4.4.2

Proof. For a fixed (x, t), we can formulate problem (4.7) as the following linear functional optimization problem.

$$\max_{\mathbb{P} \geq 0} \int_S Q(x, t, q, d^{\text{A}}, d^{\text{I}}, u) \, d\mathbb{P} \tag{4.19a}$$

$$\text{s.t.} \int_S d_i^{\text{A}} \, d\mathbb{P} = \mu_i^{\text{A}} \qquad \forall i = 1, \dots, P \tag{4.19b}$$

$$\int_S d_i^{\text{I}} \, d\mathbb{P} = \mu_i^{\text{I}} \qquad \forall i = 1, \dots, P \tag{4.19c}$$

$$\int_S u_i \, d\mathbb{P} = \mu_i^{\text{u}} \qquad \forall i = 1, \dots, P \tag{4.19d}$$

$$\int_S q_i \, d\mathbb{P} = \mu_i^{\text{q}} \qquad \forall i = 1, \dots, P \tag{4.19e}$$

$$\int_S d\mathbb{P} = 1 \tag{4.19f}$$

Letting $\rho = [\rho_1, \ldots, \rho_P]^\top$, $\alpha = [\alpha_1, \ldots, \alpha_P]^\top$, $\lambda = [\lambda_1, \ldots, \lambda_P]^\top$, $\gamma = [\gamma_1, \ldots, \gamma_P]^\top$, and θ be the dual variable associated with constraints (4.19b), (4.19c), (4.19d), and (4.19f), respectively, we present problem (4.19) in its dual form:

$$\min_{(\rho,\alpha,\lambda,\gamma)\in\mathbb{R}^P,\theta\in\mathbb{R}} \quad \sum_{i=1}^{P} \mu_i^{\text{A}}\rho_i + \mu_i^{\text{I}}\alpha_i + \mu_i^{\text{u}}\lambda_i + \mu_i^{\text{q}}\gamma_i + \theta \tag{4.20a}$$

$$\text{s.t.} \quad \sum_{i=1}^{P}(d_i^{\text{A}}\rho_i + d_i^{\text{I}}\alpha_i + u_i\lambda_i + q_i\gamma_i) + \theta \geq Q(x,t,q,d^{\text{A}},d^{\text{I}},u) \quad \forall (q,d^{\text{A}},d^{\text{I}},u) \in S \tag{4.20b}$$

where ρ, α, λ, γ, and θ are unrestricted in sign, and constraint (4.20b) is associated with the primal variable $\mathbb{P}$. Assuming that (1) $\mu_i^{\text{A}}(\mu_i^{\text{I}})$ lies in the interior of the set $\{\int_S d_i^{\text{A}}(d_i^{\text{I}})\, d\mathbb{Q} : \mathbb{Q}$ is a probability distribution over $S\}$, (2) μ_i^{q} lies in the interior of the set $\{\int_S q_i\, d\mathbb{Q} : \mathbb{Q}$ is a probability distribution over $S\}$, and (3) μ_i^{u} lies in the interior of the set $\{\int_S u_i\, d\mathbb{Q} : \mathbb{Q}$ is a probability distribution over $S\}$ for each appointment i, strong duality hold between (4.19) and (4.20) (see *Bertsimas and Popescu* (2005) for a detailed discussion on this assumption and *Jiang et al.* (2017); *Mak et al.* (2014) for applications). Observe that for fixed $(\rho,\alpha,\lambda,\gamma,\theta)$, constraint (4.20b) is equivalent to $\theta \geq \max\limits_{(q,d^{\text{A}},d^{\text{I}},u)\in S}\{Q(x,t,q,d^{\text{A}},d^{\text{I}},u) - \sum_{i=1}^{p}(d_i^{\text{A}}\rho_i + d_i^{\text{I}}\alpha_i + u_i\lambda_i + q_i\gamma_i)\}$. Since we are minimizing θ in (4.20) and $Q(x,t,\xi) \equiv \max_{y\in Y} \ \sum_{i=1}^{P}(t_i + u_i + q_i d_i^{\text{A}} + (1-q_i)d_i^{\text{I}})y_{i+1} - \sum_{i=1}^{P}(t_i+u_i)y_i - \mathcal{L}y_{P+1}$ (see Proposition 4.4.1), the dual formulation of (4.19) is equivalent to:

$$\min_{\rho,\alpha,\lambda,\gamma}\left\{\sum_{i=1}^{P} \mu_i^{\text{A}}\rho_i + \mu_i^{\text{I}}\alpha_i + \mu_i^{\text{u}}\lambda_i + \mu_i^{\text{q}}\gamma_i + \max_{y\in Y} h(x,t,y,\rho,\alpha,\lambda,\gamma)\right\} \tag{4.21}$$

where $h(x,t,y,\rho,\alpha,\lambda,\gamma) := \max\limits_{(q,d^{\text{A}},d^{\text{I}},u)\in S}\Big\{\sum_{i=1}^{P}(t_i+u_i+q_i d_i^{\text{A}}+(1-q_i)d_i^{\text{I}})y_{i+1} - \sum_{i=1}^{P}(t_i+u_i)y_i - \mathcal{L}y_{P+1} - \sum_{i=1}^{P}(d_i^{\text{A}}\rho_i + d_i^{\text{I}}\alpha_i + u_i\lambda_i + q_i\gamma_i)\Big\}$. □

4.7.5 Proof of Proposition 4.4.3

Proof. Note that the optimization problem defining function $h(x, t, y, \rho, \alpha, \lambda, \gamma)$ is separable by each appointment–i.e.,

$$\begin{aligned}
h(x, t, y, \rho, \alpha, \lambda, \gamma) = & \sum_{i=1}^{P} \max_{u_i \in [u_i^{\mathrm{L}}, u_i^{\mathrm{U}}]} u_i(y_{i+1} - y_i) + \sum_{i=1}^{P} \max_{\substack{q_i \in \{0,1\}, d_i^{\mathrm{A}} \in [d_i^{\mathrm{AL}}, d_i^{\mathrm{AU}}] \\ d_i^{\mathrm{I}} \in [d_i^{\mathrm{IL}}, d_i^{\mathrm{IU}}]}} \Big\{ q_i d_i^{\mathrm{A}} + (1 - q_i) d_i^{\mathrm{I}} \Big\} y_{i+1} \\
& + \sum_{i=1}^{P} (t_i y_{i+1} - t_i y_i) - \mathcal{L} y_{P+1} + \sum_{i=1}^{P} \Big(\max_{d_i^{\mathrm{A}} \in [d_i^{\mathrm{AL}}, d_i^{\mathrm{AU}}]} -d_i^{\mathrm{A}} \rho_i + \max_{d_i^{\mathrm{I}} \in [d_i^{\mathrm{IL}}, d_i^{\mathrm{IU}}]} -d_i^{\mathrm{I}} \alpha_i \\
& + \max_{u_i \in [u_i^L, u_i^U]} -u_i \lambda_i + \max_{q_i \in \{0,1\}} -q_i \gamma_i \Big) \\
= & (-t_1 - u_1^{\mathrm{L}}) y_1 + \sum_{i=1}^{P} (u_i^{\mathrm{L}} - u_i^{\mathrm{U}})(y_{i+1} - y_i)^+ + \sum_{i=2}^{P} (-t_i - u_i^{\mathrm{L}} + t_{i-1} + u_{i-1}^{\mathrm{L}}) y_i \\
& + \sum_{i=2}^{P} \min\{d_{i-1}^{\mathrm{AL}}, d_{i-1}^{\mathrm{IL}}\} y_i + \Big[\max\{d_{i-1}^{\mathrm{AU}}, d_{i-1}^{\mathrm{IU}}\} - \min\{d_{i-1}^{\mathrm{AL}}, d_{i-1}^{\mathrm{IL}}\} \Big] (y_i)^+ + \Big(t_P \\
& + u_P^L - \mathcal{L} + \min\{d_P^{\mathrm{AL}}, d_P^{\mathrm{IL}}\} \Big) y_{P+1} + \Big(\max\{d_P^{\mathrm{AU}}, d_P^{\mathrm{IU}}\} - \min\{d_P^{\mathrm{AL}}, d_P^{\mathrm{IL}}\} \Big) (y_{P+1})^+ \\
& + \sum_{i=1}^{P} K_i
\end{aligned} \tag{4.22}$$

where $(a)^+ = \max(a, 0)$ and $K_i = -\Big(d_i^{\mathrm{AU}} \rho_i + (d_i^{\mathrm{AL}} - d_i^{\mathrm{AU}})(\rho_i)^+ \Big) - \Big(d_i^{\mathrm{IU}} \alpha_i + (d_i^{\mathrm{IL}} - d_i^{\mathrm{IU}})(\alpha_i)^+ \Big) - \Big(u_i^{\mathrm{U}} \lambda_i + (u_i^{\mathrm{L}} - u_i^{\mathrm{U}})(\lambda_i)^+ \Big) + (-\gamma_i)^+$ for all $i = 1, \ldots, P$. It is straightforward to see that function $h(x, t, y, \rho, \alpha, \lambda, \gamma)$ is convex in variables y. Hence, $\max_{y \in Y} h(x, t, y, \rho, \alpha, \lambda, \gamma)$ is a convex maximization problem. It follows from the fundamental convex analysis (see, e.g., *Boyd and Vandenberghe* (2004)) that there exists an optimal solution y^* to $\max\limits_{y \in Y} h(x, t, y, \rho, \alpha, \lambda, \gamma)$ at one of the extreme point of the polyhedron Y defined in (4.9).

It is easy to see from the definition of Y in (4.9) that any any extreme point $\hat{y}$ of Y satisfy (*i*) $\hat{y}_{P+2} = -c_{P+1}^{\mathrm{g}} = 0$, and (*ii*) for all $i = 1, \ldots, P+1$, the dual constraint $\hat{y}_{i+1} + c_i^{\mathrm{w}} \geq \hat{y}_i \geq -c_i^{\mathrm{g}}$ is binding at either the lower bound or upper bound. This defines a unique one-to-one correspondence between an extreme point $\hat{y}$ of Y and a partition of integers 1,..., $P + 2$ into intervals. Each interval

$\{k, ..., j\} \subseteq \{1, ..., P+2\}$ in the partition has the following property; $y_j = -c_j^{\mathrm{g}}$ (lower bound value) and other $y_i = y_{i+1} + c_i^{\mathrm{w}}, \forall i = k, ..., j-1$ (upper bound value). As a result of this property, for each interval $\{k, ..., j\}$ in the partition and $i \in \{k, \dots, j\}$, we recursively obtain the value of y_i as follow:

$$y_i = \pi_{i,j} = -c_j^g + \sum_{\ell=i}^{j-1} c_\ell^{\mathrm{w}} \quad \forall 1 \leq i \leq j \leq P+2 \text{ and } y_{P+2} = \pi_{P+2,P+2} = 0 \tag{4.23}$$

Thus, the problem of finding an optimal extreme point $y \in Y$ (i.e., $\max\limits_{y \in Y} \; h(x, t, y, \rho, \alpha, \lambda, \gamma)$) can be transformed into finding an optimal partition of the integers into intervals as follows. We define a binary variable $b_{kj} = 1$ if and only if the interval $\{k, ..., j\}$ belong to the partition and $b_{kj} = 0$ otherwise for all $1 \leq k \leq j \leq P+2$. b_{kj} represent a valid partition if and only if each index *i* belong to to exactly one interval, i.e., $\sum_{k=1}^{i} \sum_{j=i}^{P+2} b_{kj} = 1$, for all $i = 1, ..., P+2$. Using binary variables b_{kj} and $h(x, t, y, \rho, \alpha, \lambda, \gamma)$ defintion in (4.22), we reformulate the maximization problem $\max\limits_{y \in Y} \; h(x, t, y, \rho, \alpha, \lambda, \gamma)$ in (4.11) as:

$$\begin{aligned}
\max_b \; & \sum_{j=1}^{P+2} (-t_1 - u_1^{\mathrm{L}}) \pi_{1,j} b_{1,j} + \sum_{i=1}^{P} \sum_{j=i+1}^{P+2} (u_i^{\mathrm{U}} - u_i^{\mathrm{L}}) (\pi_{i+1,j} + c_i^g)^+ b_{i+1,j} \\
& + \sum_{i=2}^{P} \sum_{k=1}^{i} \sum_{j=i}^{P+2} \big(-t_i - u_i^{\mathrm{L}} + t_{i-1} + u_{i-1}^{\mathrm{L}} + \min\{d_{i-1}^{\mathrm{AL}}, d_{i-1}^{\mathrm{IL}}\} \big) \pi_{i,j} b_{k,j} \\
& + \sum_{i=2}^{P} \sum_{k=1}^{i} \sum_{j=i}^{P+2} \big(\max\{d_{i-1}^{\mathrm{AU}}, d_{i-1}^{\mathrm{IU}}\} - \min\{d_{i-1}^{\mathrm{AL}}, d_{i-1}^{\mathrm{IL}}\} \big) (\pi_{i,j})^+ b_{k,j} \\
& + \sum_{k=1}^{P+1} \sum_{j=P+1}^{P+2} \big(t_P + u_P^L - \mathcal{L} + \min\{d_P^{\mathrm{AL}}, d_P^{\mathrm{IL}}\} \big) \pi_{P+1,j} b_{k,j} \\
& + \sum_{k=1}^{P+1} \sum_{j=P+1}^{P+2} (\max\{d_P^{\mathrm{AU}}, d_P^{\mathrm{IU}}\} - \min\{d_P^{\mathrm{AL}}, d_P^{\mathrm{IL}}\} \big) (\pi_{P+1,j})^+ b_{k,j} \\
& + \sum_{i=1}^{P} \sum_{k=1}^{i} \sum_{j=i}^{P+2} K_i b_{k,j}
\end{aligned} \tag{4.24a}$$

$$\text{s.t.} \quad \sum_{k=1}^{i}\sum_{j=i}^{P+2} b_{kj} = 1, \qquad \text{for } i = 1, ..., P+2 \tag{4.24b}$$

$$b_{kj} \in \{0,1\}, \qquad \text{for } 1 \le k \le j \le P+2 \tag{4.24c}$$

With $t_{P+1} = u^{\text{U, L}}_{P+1} = d^{\text{AU,AL}}_{P+1} = d^{\text{IU,IL}}_{P+1} = t_{P+2} = u^{\text{U,L}}_{P+2} = d^{\text{AU,AL}}_{P+2} = d^{\text{U,L}}_{P+2} = \pi_{P+2,P+2} = 0$. The term $(u^{\text{U}}_i - u^{\text{L}}_i)(\pi_{i+1,j} + c^g_i)^+$ in (4.24a) is equivalent to $(u^{\text{U}}_i - u^{\text{L}}_i)(y_{i+1} - y_i)^+$ in (4.22) because $(y_{i+1} - y_i)^+$ is positive if and only if i and $i+1$ belong to different partitions. That is if i and $i+1$ belong to the same partition then $y_{i+1} - y_i = -c^w_i \le 0$, and so $(y_{i+1} - y_i)^+ = (-c^w_i)^+ = 0$. If not, then $i \in \{k, \dots, i\}$ for $1 \le k \le i \le P+1$ and $i+1 \in \{i+1, \dots, j\}$ for $i+1 \le j \le P+2$. In this case, $(y_{i+1} - y_i)^+ = (y_{i+1} + c^g_i)^+ = (\pi_{i+1,j} + c^g_i)^+$ by the partition in (4.23).

Note that the constraint matrix associated with constraints (4.24b)–(4.24c) is totally unimodular. Therefore, its LP relaxation, which is obtained by replacing the binary constraints $b_{kj} \in \{0,1\}$ in (4.24c) by $0 \le b_{k,j} \le 1$, has a binary optimal solution. And so, the integer program in (4.24) has the same optimal objective value as its LP relaxation as well as the dual problem of the LP relaxation. The dual problem of the LP is given by $\min\{\sum_{i=1}^{P+2} \beta_i$: (4.13b) − (4.13d)$\}$ □

4.7.6 Proof of Proposition 4.4.4

Proof. Incorporating the sequencing decisions in (4.14) leads to the following MINLP formulation.

$$\min_{\substack{x,\in\mathcal{X},t\in\mathcal{T}\\ \rho,\alpha,\lambda,\gamma,\beta}} \sum_{i=1}^{P}\sum_{p=1}^{P} \mu^{\text{A}}_p x_{p,i}\rho_i + \mu^{\text{I}}_p x_{p,i}\alpha_i + \mu^{\text{u}}_p x_{p,i}\lambda_i + \mu^{\text{q}}_p x_{p,i}\gamma_i + \sum_{i=1}^{P+2}\beta_i \tag{4.25a}$$

$$\begin{aligned}\text{s.t.} \sum_{i=1}^{j}\beta_i \ge (-t_1 - \sum_{p=1}^{P} u^{\text{L}}_p x_{p,1})\pi_{1,j} + \sum_{i=P+1}^{\min\{j,P+1\}} \Big(t_P - \mathcal{L} + \sum_{p=1}^{P}\big(u^{\text{L}}_p + \min\{d^{\text{AL}}_p, d^{\text{IL}}_p\}\big)x_{p,P}\Big)\pi_{P+1,j} \\ + \sum_{i=2}^{\min\{j,P\}} \Big(t_{i-1} - t_i - \sum_{p=1}^{P} u^{\text{L}}_p x_{p,i} + \sum_{p=1}^{P}\big(u^{\text{L}}_p + \min\{d^{\text{AL}}_p, d^{\text{IL}}_p\}\big)x_{p,i-1}\Big)\pi_{i,j} + \sum_{i=1}^{\min\{j,P\}} K''_i\end{aligned}$$

$$+ \sum_{i=2}^{\min\{j,P+1\}} \Big(\sum_{p=1}^{P} \big(\max\{d_p^{\text{AU}}, d_p^{\text{IU}}\} - \min\{d_p^{\text{AL}}, d_p^{\text{IL}}\}\big) x_{p,i-1}\Big) (\pi_{i,j})^+, \ \forall j = 1, \dots, P+2 \tag{4.25b}$$

$$\begin{aligned}
\sum_{i=k}^{j} \beta_i \geq & \sum_{p=1}^{P} (u_p^{\text{U}} - u_p^{\text{L}}) x_{p,k-1} (\pi_{k,j} + c_{k-1}^{g})^+ + \sum_{i=\min\{k,P+1\}}^{\min\{j,P\}} K_i'' \\
& + \sum_{i=P+1}^{\min\{j,P+1\}} \Big(t_P - \mathcal{L} + \sum_{p=1}^{P} \big(u_p^{\text{L}} + \min\{d_p^{\text{AL}}, d_p^{\text{IL}}\}\big) x_{p,P}\Big) \pi_{P+1,j} \\
& + \sum_{i=\min\{k,P+1\}}^{\min\{j,P\}} \Big(t_{i-1} - t_i - \sum_{p=1}^{P} u_p^{\text{L}} x_{p,i} + \sum_{p=1}^{P} \big(u_p^{\text{L}} + \min\{d_p^{\text{AL}}, d_p^{\text{IL}}\}\big) x_{p,i-1}\Big) \pi_{i,j} \\
& + \sum_{i=\min\{k,P+1\}}^{\min\{j,P+1\}} \Big(\sum_{p=1}^{P} \big(\max\{d_p^{\text{AU}}, d_p^{\text{IU}}\} - \min\{d_p^{\text{AL}}, d_p^{\text{IL}}\}\big) x_{p,i-1}\Big) (\pi_{i,j})^+, \\
& \forall k = 2, \dots, P+1, \ j = k, \dots, P+2
\end{aligned} \tag{4.25c}$$

$$\beta_{P+2} \geq 0, z_i \geq 0, \ z_i \geq \rho_i, \ v_i \geq 0, \ v_i \geq \alpha_i, \ r_i \geq 0, \ r_i \geq \lambda_i, \ e_i \geq 0, \ e_i \geq -\gamma_i, \ \forall i \leq P \tag{4.25d}$$

where for all $i = 1, \dots, P$, $K_i'' = -\Big(\sum_{p=1}^{P} \big(d_p^{\text{AU}} x_{p,i} \rho_i + (d_p^{\text{AL}} - d_p^{\text{AU}}) x_{p,i} z_i \big) \Big) - \Big(\sum_{p=1}^{P} \big(d_p^{\text{IU}} x_{p,i} \alpha_i + (d_p^{\text{IL}} - d_p^{\text{IU}}) x_{p,i} v_i \big) \Big) - \Big(\sum_{p=1}^{P} \big(u_p^{\text{U}} x_{p,i} \lambda_i + (u_p^{\text{L}} - u_p^{\text{U}}) x_{p,i} r_i \big) \Big) + e_i$. Note that the objective function (4.25a) and the term K_i'' in constraints (4.25b)–(4.25c) contain the bilinear terms $x_{p,i}\rho_i, \ x_{p,i}\alpha_i, \ x_{p,i}\lambda_i$ and $x_{p,i}\gamma_i$ with binary variables $x_{p,i}$ and continuous variables $\rho_i, \ \alpha_i, \ \lambda_i$ and γ_i. Additionally, K_i'' contains the bilinear terms $x_{p,i}z_i$, $x_{p,i}v_i$, and $x_{p,i}r_i$. To linearize this MINLP formulation, we define $\eta_{p,i} = x_{p,i}\rho_i$, $\tau_{p,i} = x_{p,i}\alpha_{p,i}$, $\Lambda_{p,i} = x_{p,i}\lambda_i$, $\Gamma_i = x_{p,i}\gamma_i$, $\zeta_{p,i} = x_{p,i}z_i$, $\nu_{p,i} = x_{p,i}v_i$, and $\varphi_{p,i} = x_{p,i}r_i$. We also introduce the following McCormick inequalities (4.26a)–(4.26g) for variables $\zeta_{p,i}$, $\nu_{p,i}$, $\varphi_{p,i}$, $\eta_{p,i}$, $\tau_{p,i}$, $\Lambda_{p,i}$, and Γ_i respectively:

$$\zeta_{p,i} \geq 0, \ \ \zeta_{p,i} \geq z_i - (1 - x_{p,i})\overline{z}_i, \ \ \zeta_{p,i} \leq z_i, \ \ \zeta_{p,i} \leq x_{p,i}\overline{z}_i \tag{4.26a}$$

$$\nu_{p,i} \geq 0, \ \ \nu_{p,i} \geq v_i - (1 - x_{p,i})\overline{v}_i, \ \ \nu_{p,i} \leq v_i, \ \ \nu_{p,i} \leq x_{p,i}\overline{v}_i \tag{4.26b}$$

$$\varphi_{p,i} \geq 0, \;\; \varphi_{p,i} \geq r_i - (1 - x_{p,i})\overline{r}_i, \;\; \varphi_{p,i} \leq r_i, \;\; \varphi_{p,i} \leq x_{p,i}\overline{r}_i \tag{4.26c}$$

$$\eta_{p,i} \geq x_{p,i}\underline{\rho}_i, \;\; \eta_{p,i} \geq \rho_i - (1 - x_{p,i})\overline{\rho}_i, \;\; \eta_{p,i} \leq x_{p,i}\overline{\rho}_i, \;\; \eta_{p,i} \leq \rho_i - (1 - x_{p,i})\underline{\rho}_i \tag{4.26d}$$

$$\tau_{p,i} \geq x_{p,i}\underline{\alpha}_i, \;\; \tau_{p,i} \geq \alpha_i - (1 - x_{p,i})\overline{\alpha}_i, \;\; \tau_{p,i} \leq x_{p,i}\overline{\alpha}_i, \;\; \tau_{p,i} \leq \alpha_i - (1 - x_{p,i})\underline{\alpha}_i \tag{4.26e}$$

$$\Lambda_{p,i} \geq x_{p,i}\underline{\lambda}_i, \;\; \Lambda_{p,i} \geq \lambda_i - (1 - x_{p,i})\overline{\lambda}_i, \;\; \Lambda_{p,i} \leq x_{p,i}\overline{\lambda}_i, \;\; \Lambda_{p,i} \leq \lambda_i - (1 - x_{p,i})\underline{\lambda}_i \tag{4.26f}$$

$$\Gamma_{p,i} \geq x_{p,i}\underline{\gamma}_i, \;\; \Gamma_{p,i} \geq \gamma_i - (1 - x_{p,i})\overline{\gamma}_i, \;\; \Gamma_{p,i} \leq x_{p,i}\overline{\gamma}_i, \;\; \Gamma_{p,i} \leq \gamma_i - (1 - x_{p,i})\underline{\gamma}_i \tag{4.26g}$$

where coefficients $(\underline{z}, \underline{v}, \underline{r}, \underline{\rho}, \underline{\alpha}, \underline{\lambda}, \underline{\gamma})$ and $(\overline{z}, \overline{v}, \overline{r}, \overline{\rho}, \overline{\alpha}, \overline{\lambda}, \overline{\gamma})$ are respectively valid lower and upper bounds on the values of variables $(z, v, r, \rho, \alpha, \lambda, \gamma)$. Letting $K'' = -\Big(\sum_{p=1}^{P} d_p^{\text{AU}}\eta_{p,i} + \sum_{p=1}^{P}(d_p^{\text{AL}} - d_p^{\text{AU}})\zeta_{p,i}\Big) - \Big(\sum_{p=1}^{P} d_p^{\text{IU}}\tau_{p,i} - \sum_{p=1}^{P}(d_p^{\text{IL}} - d_p^{\text{IU}})\nu_{p,i}\Big) - \Big(\sum_{p=1}^{P} u_p^{\text{U}}\Lambda_{p,i} + \sum_{p=1}^{P}(u_p^{\text{L}} - u_p^{\text{U}})\varphi_{p,i}\Big) + e_i$ for all $i \in [P]$, formulation (4.25) (equivalently, the DR scheduling model in (4.6)) is equivalent to the following MILP.

$$\text{(DR-bimodal)} \quad \min \; \sum_{i=1}^{P}\sum_{p=1}^{P} \mu_p^{\text{A}}\eta_{p,i} + \mu_p^{\text{I}}\tau_{p,i} + \mu_p^{\text{u}}\nu_{p,i} + \mu_p^{\text{q}}\Gamma_{p,i} + \sum_{i=1}^{P+2}\beta_i \tag{4.27a}$$

$$\text{s.t.} \;\; (x \in \mathcal{X}, t \in \mathcal{T}) \tag{4.27b}$$

$$\text{(4.25b) – (4.25d), (4.26d) – (4.26g)} \tag{4.27c}$$

□

4.7.7 Strengthening the MILP Formulation of DROCS

In this section, we derive tight lower and upper bounds on the values of variables $(\rho, \alpha, \lambda, \gamma, z, v, r)$ in McCormick inequalities (4.26) of the MILP formulation in (4.15). First, in Lemma 4.7.1 we derive tight upper and lower bounds on variables $y \in Y$ defined in (4.9). Then, in Propositions 4.7.2–4.7.5, we use the results of Lemma 4.7.1, to derive tight bounds on variables $(\rho, \alpha, \lambda, \gamma, z, v, r)$.

Lemma 4.7.1. *$\underline{y}_i = -c_i^{g}$ and $\overline{y}_i = \sum\limits_{j=i}^{P+1} c_j^{w}$ are respectively valid lower and upper bounds on* *variable y_i in* (4.11)*, for $i = 1, \ldots, P+1$.*

Proof. It's straightforward to derive $\underline{y}_i = -c^{\text{g}}_i$ and $\overline{y}_i = \sum\limits_{j=i}^{P+1} c^{\text{w}}_j$ from the the definition of polyhedron $Y := \{y_{P+2} = 0,\ c^{\text{w}}_i + y_{i+1} \geq y_i \geq -c^{\text{g}}_i \text{ for } i = 1, ..., P+1\}$. □

Proposition 4.7.2. *$\underline{y}_{i+1}$ and $\overline{y}_{i+1}$ are respectively valid lower and upper bounds on variables (ρ_i, α_i), for $i = 1, \ldots, P$.*

Proof. First, we prove that $\rho_i \in [\underline{y}_{i+1}, \overline{y}_{i+1}]$. Observe from the objective of the DROCS problem in (4.11) that variables ρ_i are multiplied by parameters μ^A_i and variables d^A_i for all $i = 1, \ldots, P$. And so, for fixed $y_{i+1} \in Y$, the joint contribution of ρ_i and d^A_i, $\forall i = 1, \ldots, P$, to the objective of problem (4.11) equals:

$$\mu^A_i \rho_i + \max_{d^A_i \in [d^{AL}_i, d^{AU}_i]} (q_i y_{i+1} - \rho_i) d^A_i = \mu^A_i \rho_i + (q_i y_{i+1} - \rho_i) d^{AL}_i + (d^{AU}_i - d^{AL}_i)(q_i y_{i+1} - \rho_i)^+. \tag{4.28}$$

Suppose that $\rho_i > \overline{y}_{i+1}$. In this case, $(q_i y_{i+1} - \rho_i)^+ = 0$ and so ρ_i contribute to the objective value of problem (4.11) by $(\mu^A_i - d^{AL}_i)\rho_i$. Let $\rho'_i = \rho_i - \epsilon$ with $\epsilon > 0$. Since $(\mu^A_i - d^{AL}_i) \geq 0$, then $(\mu^A_i - d^{AL}_i)\rho'_i < (\mu^A_i - d^{AL}_i)\rho_i$, i.e., ρ'_i improves the objective value of problem (4.11). It follows that, without any loss of optimality, $\overline{\rho}_i = \overline{y}_{i+1}$ is a valid upper bound on ρ_i for $i = 1, \ldots, P$.

Conversely, suppose that $\rho_i < \underline{y}_{i+1}$. In this case, ρ_i contribute to the objective value of problem (4.11) by $(\mu^A_i - d^{AU}_i)\rho_i$. Let $\rho'_i = \rho_i + \epsilon$ with $\epsilon > 0$. Since $(\mu^A_i - d^{AU}_i) \leq 0$, then $(\mu^A_i - d^{AU}_i)\rho'_i < (\mu^A_i - d^{AU}_i)\rho_i$, i.e., ρ'_i improves the objective value of problem (4.11). It follows that, without any loss of optimality, $\underline{\rho}_i = \underline{y}_{i+1}$ is a valid lower bound on ρ_i for $i = 1, \ldots, P$.

Second, we observe from (4.11) that variables α_i are multiplied by parameters μ^I_i and variables d^I_i for $i = 1, \ldots, P$. And so, for fixed y_{i+1}, the joint contribution of α_i and d^I_i, $\forall i = 1, \ldots, P$, to the objective of problem (4.11) equals:

$$\mu_i^I\alpha_i + \max_{d_i^I \in [d_i^{IL}, d_i^{IU}]} ((1-q_i)y_{i+1} - \alpha_i)d_i^I = \mu_i^I + ((1-q_i)y_{i+1} - \alpha_i)d_i^{IL}\alpha_i \tag{4.29}$$

$$+ (d_i^{IU} - d_i^{IL})((1-q_i)y_{i+1} - \alpha_i)^+. \tag{4.30}$$

It follows that $\alpha_i \in [\underline{y}_{i+1}, \overline{y}_{i+1}]$, for $i = 1, \ldots, P$. □

Proposition 4.7.3. *$\underline{\lambda}_i = \underline{y}_{i+1} - \overline{y}_i$ and $\overline{\lambda}_i = \overline{y}_{i+1} - \underline{y}_i$ are respectively valid lower and upper bounds on variables λ_i for $i = 1, \ldots, P$.* (4.11)

Proof. Observe from the objective of the DROCS problem in (4.11) that variables λ_i are multiplied by parameters μ_i^u and variables u_i for $i = 1, \ldots, P$. And so, for fixed y_{i+1} and y_i, the joint contribution of variables λ_i and u_i, for all $i = 1, \ldots, P$, to the objective of problem (4.11) equals:

$$\max_{u_i \in [u_i^L, u_i^U]} (y_{i+1} - y_i - \lambda_i)u_i = (y_{i+1} - y_i - \lambda_i)u_i^L + (y_{i+1} - y_i - \lambda_i)^+(u_i^U - u_i^L) \tag{4.31}$$

Using the same proof techniques of proposition 4.7.2, one can show that $\lambda_i \in [\underline{y}_{i+1} - \overline{y}_i, \overline{y}_{i+1} - \underline{y}_i]$, for $i = 1, \ldots, P$. □

Proposition 4.7.4. *$\underline{\gamma}_i = -M\overline{y}_{i+1}$ and $\overline{\gamma}_i = M\overline{y}_{i+1}$ are respectively valid lower and upper bounds on variables γ_i for $i = 1, \ldots, P$. Where,*

M=max $\left\{|d_i^{AU} - d_i^{IU}|,\ |d_i^{AU} - d_i^{IL}|,\ |d_i^{AL} - d_i^{IU}|,\ |d_i^{AL} - d_i^{IL}|\right\}$.

Proof. Observe from the objective of the DROCS problem in (4.11) that variables γ_i are only multiplied by parameters μ_i^q and variables q_i for $i = 1, \ldots, P$. And so, for fixed y_{i+1}, the joint contribution of variables γ_i and q_i to the objective of problem (4.11) equals:

$$\mu_i^q\gamma_i + \max_{\substack{q_i \in \{0,1\} \\ d_i^A \in [d_i^{AL}, d_i^{AU}] \\ d_i^I \in [d_i^{IL}, d_i^{IU}]}} ((d_i^A - d_i^I)y_{i+1} - \gamma_i)q_i = \mu_i^q\gamma_i + (M|y_{i+1}| - \gamma_i)^+. \tag{4.32}$$

Using the same proof techniques of proposition 4.7.2, one can show that $\gamma_i \in [-M\overline{y}_{i+1}, M\overline{y}_{i+1}]$, for $i = 1, \ldots, P$.

□

Proposition 4.7.5. $(\overline{z}_i = \overline{\rho}_i, \overline{v}_i = \overline{\alpha}_i, \overline{r}_i = \overline{\lambda}_i)$ *are valid upper bounds on variables (*z_i, v_i, r_i*), for all* $i = 1, \ldots, P$.

Proof. By constraints (4.25d) and the objective of minimizing $(\rho_i, \alpha_i, \lambda_i, \beta_i)$, $z_i = \max(\rho_i, 0)$, $v_i = \max(\alpha_i, 0)$, and $r_i = \max(\lambda_i, 0)$. It follows, that $z_i \in [0, \overline{\rho}_i]$, $v_i \in [0, \overline{\alpha}_i]$, and $r_i \in [0, \overline{\lambda}_i]$, for $i = 1, \ldots, P$. □

4.7.8 Stochastic Mixed-Integer Linear Program (SMILP)

The following SMILP-bimodal minimize the expected weighted sum of the scheduling metrics via the sample average approximation (SAA) approach with N scenarios (see, e.g., *Kim et al.* (2015); *Kleywegt et al.* (2002) and references therein for detailed information on SAA).

$$\min_{x\in\mathcal{X},t\in\mathcal{T},w,a,g} \ \frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{P}(c_i^{\mathrm{w}}w_i^n + c_i^{\mathrm{g}}g_i^n) + c^{\mathrm{o}}w_{P+1}^n \tag{4.33a}$$

$$\text{s.t.} \quad a_i^n = t_i + \sum_{p=1}^{P} u_p^n x_{p,i}, \quad \forall i \in [P], \forall n \in [N] \tag{4.33b}$$

$$w_1^n - g_1^n + a_1^n = 0, \ \ \forall n \in [N] \tag{4.33c}$$

$$w_i^n - g_i^n = a_{i-1}^n + w_{i-1}^n \sum_{p=1}^{P}\left(q_p^n d_p^{An} + (1-q_p^n)d_p^{In}\right)x_{p,i-1} - a_i^n, \ \ \forall i \in [2,P]_{\mathbb{Z}}, \forall n \in [N] \tag{4.33d}$$

$$w_{P+1}^n - g_{P+1}^n = a_P^n + w_P^n \sum_{p=1}^{P}\left(q_p^n d_p^{An} + (1-q_p^n)d_p^{In}\right)x_{p,P} - \mathcal{L}, \ \ \forall n \in [N] \tag{4.33e}$$

$$(w_i^n, g_i^n) \geq 0, \ \forall i \in [P+1], \ \forall n \in [N] \tag{4.33f}$$

where $q_p^n, d_p^{\mathrm{A}n}, d_p^{\mathrm{I}p}$, and u_p^n are respectively realizations of parameters q_p, d_p^{A}, d_p^{I}, and u_p of appointment *p* in scenario n from their known distributions for all $p \in [P]$ and $n \in [N]$. Variables w_i^n, g_i^n, and w_{P+1}^n are the recourse waiting time of appointment i, provider idle time after appointment i, and the clinic overtime in scenario *n*, respectively, for all $n \in [N]$. Constraints (4.33c)–(4.33f) computes the (waiting time, idle time, and overtime) values based on the schedule (x, t) and a scenario n of parameters ($q_p^n, d_p^{\mathrm{A}n}, d_p^{\mathrm{I}n}, u_p^n$) for all $n \in [N]$.

In the SMILP-plain, we assume that colonoscopy durations follow one probability distribution, and so we replace the term $q_p^n d_p^{An} + (1-q_p^n)d_p^{In}$ with d_p^n in constraints (4.33d) and (4.33e), where d_p^n represent colonoscopy duration in scenario n.

4.7.9 DR-plain Formulation

In this section, we present the DR-plain model in which we ignore prep adequacy and assume that colonoscopy durations, d, follow a single (unknown) probability distribution with a known mean μ^{d} and support $S^d := \{d \geq 0 : d^L \leq d \leq d^U\}$. And so, we assume that $\mathbb{P}$ of $\xi := [d, u]^\top$ belongs to an ambiguity set $\mathcal{F}^{\text{plain}}(S, \mu)$ of possible distributions, which incorporate the known support $S = S^{\text{d}} \times S^{\text{u}}$ and mean $\mu := \mathbb{E}_{\mathbb{P}}[\xi] = [\mu^{\text{d}}, \mu^{\text{u}}]^\top$. Using ambiguity set $\mathcal{F}^{\text{plain}}(S, \mu) := \big\{\mathbb{P} \in \mathcal{P}(S) : \int_S d\mathbb{P} = 1, \mathbb{E}_{\mathbb{P}}[\xi] = \mu\big\}$, we formulate the DR-plain as the following min-max problem:

$$\text{(Dr-plain)} \ \min_{x,t} \sup_{\mathbb{P} \in \mathcal{F}^{\text{plain}}(S,\mu)} \mathbb{E}_{\mathbb{P}}[Q(x, t, \xi)] \tag{4.34}$$

where $\xi := [d, u]^\top$. We follow the same logic in Section 4.4.5 to derive the following MILP formulation of DR-plain in (4.34).

$$\min \ \sum_{i=1}^{P} \sum_{p=1}^{P} \mu_p^{\text{d}} \eta_{p,i} + \mu_p^{\text{u}} \Lambda_{p,i} + \sum_{i=1}^{P+2} \beta_i \tag{4.35a}$$

$$\begin{aligned}
\text{s.t.} \ \sum_{i=1}^{j} \beta_i \geq{} & (-t_1 - \sum_{p=1}^{P} u_p^{\text{L}} x_{p,1}) \pi_{1,j} + \sum_{i=2}^{\min\{j,P\}} \Big(t_{i-1} - t_i - \sum_{p=1}^{P} u_p^{\text{L}} x_{p,i} + \sum_{p=1}^{P} \big(u_p^{\text{L}} + d_p^{\text{L}}\big) x_{p,i-1}\Big) \pi_{i,j} \\
& + \sum_{i=2}^{\min\{j,P+1\}} \Big(\sum_{p=1}^{P} \big(d_p^{\text{U}} - d_p^{\text{L}}\big) x_{p,i-1}\Big) (\pi_{i,j})^+ + \sum_{i=1}^{\min\{j,P\}} V_i \\
& + \sum_{i=P+1}^{\min\{j,P+1\}} \Big(t_P - \mathcal{L} + \sum_{p=1}^{P} \big(u_p^{\text{L}} + d_p^{\text{L}}\big) x_{p,P}\Big) \pi_{P+1,j}, \quad \forall j \in [1, P+2]_{\mathbb{Z}}
\end{aligned} \tag{4.35b}$$

$$\begin{aligned}
\sum_{i=k}^{j} \beta_i \geq{} & \sum_{p=1}^{P} (u_p^{\text{U}} - u_p^{\text{L}}) x_{p,k-1} (\pi_{k,j} + c_{k-1}^g)^+ + \sum_{i=\min\{k,P+1\}}^{\min\{j,P+1\}} \Big(\sum_{p=1}^{P} \big(d_p^{\text{U}} - d_p^{\text{L}}\big) x_{p,i-1}\Big) (\pi_{i,j})^+ \\
& + \sum_{i=\min\{k,P\}}^{\min\{j,P\}} \Big(t_{i-1} - t_i - \sum_{p=1}^{P} u_p^{\text{L}} x_{p,i} + \sum_{p=1}^{P} \big(u_p^{\text{L}} + d_p^{\text{L}}\big) x_{p,i-1}\Big) \pi_{i,j} + \sum_{i=\min\{k,P\}}^{\min\{j,P\}} V_i
\end{aligned}$$

$$+ \sum_{i=P+1}^{\min\{j,P+1\}} \Big(t_P - \mathcal{L} + \sum_{p=1}^{P} \big(u_p^{\mathrm{L}} + d_p^{\mathrm{L}}\big) x_{p,P}\Big) \pi_{P+1,j}, k \in [2, P+1]_{\mathbb{Z}}, j \in [k, P+2] \tag{4.35c}$$

$$\sum_{i=P+1}^{j} \beta_i \geq \sum_{p=1}^{P} (u_p^{\mathrm{U}} - u_p^{\mathrm{L}}) x_{p,P} (\pi_{P+1,j} + c_P^g)^+ + \Big(t_P - \mathcal{L} + \sum_{p=1}^{P} \big(u_p^{\mathrm{L}} + d_p^{\mathrm{L}}\big) x_{p,P}\Big) \pi_{P+1,j}$$
$$+ \Big(\sum_{p=1}^{P} \big(d_p^{\mathrm{U}} - d_p^{\mathrm{L}}\big) x_{p,P} \Big) \, (\pi_{P+1,j})^+, \;\; \forall j = P+1, P+2 \tag{4.35d}$$

$$\eta_{p,i} \geq x_{p,i} \underline{\rho}_i, \; \eta_{p,i} \geq \rho_i - (1 - x_{p,i}) \overline{\rho}_i, \; \forall (p,i) \in [P] \tag{4.35e}$$

$$\eta_{p,i} \leq x_{p,i} \overline{\rho}_i, \; \eta_{p,i} \leq \rho_i - (1 - x_{p,i}) \underline{\rho}_i, \; \forall (p,i) \in [P] \tag{4.35f}$$

$$\Lambda_{p,i} \geq x_{p,i} \underline{\lambda}_i, \; \Lambda_{p,i} \geq \lambda_i - (1 - x_{p,i}) \overline{\lambda}_i, \quad \forall (p,i) \in [P] \tag{4.35g}$$

$$\Lambda_{p,i} \leq x_{p,i} \overline{\lambda}_i, \; \Lambda_{p,i} \leq \lambda_i - (1 - x_{p,i}) \underline{\lambda}_i, \; \forall (p,i) \in [P] \tag{4.35h}$$

$$\zeta_{p,i} \geq 0, \; \zeta_{p,i} \geq z_i - (1 - x_{p,i}) \overline{z}_i, \;\; \zeta_{p,i} \leq z_i, \; \zeta_{p,i} \leq x_{p,i} \overline{z}_i, \; \forall (p,i) \in [P] \tag{4.35i}$$

$$\varphi_{p,i} \geq 0, \; \varphi_{p,i} \geq r_i - (1 - x_{p,i}) \overline{r}_i, \;\; \varphi_{p,i} \leq r_i, \; \varphi_{p,i} \leq x_{p,i} \overline{r}_i, \forall (p,i) \in [P] \tag{4.35j}$$

$$\beta_{P+2} \geq 0, z_i \geq 0, z_i \geq \rho_i, v_i \geq 0, v_i \geq \alpha_i, r_i \geq 0, \; r_i \geq \lambda_i, e_i \geq 0, e_i \geq -\gamma_i, \forall i \in [P] \tag{4.35k}$$

where $V_i = -\Big(\sum_{p=1}^{P} d_p^{\mathrm{U}} \eta_{p,i} + \sum_{p=1}^{P} (d_p^{\mathrm{L}} - d_p^{\mathrm{U}}) \zeta_{p,i} \Big) - \Big(\sum_{p=1}^{P} u_p^{\mathrm{U}} \Lambda_{p,i} + \sum_{p=1}^{P} (u_p^{\mathrm{L}} - u_p^{\mathrm{U}}) \varphi_{p,i} \Big) + e_i$ for all $i = 1, \ldots, P$.

4.7.10 Random Sequences

Figure 4.7 presents the random sequencing matrices $\mathbf{X}_1, \ldots, \mathbf{X}_{1000}$, i.e., the assignment of procedures to positions in each of 1000 randomly generated sequences. Figure 4.8 presents a histogram of procedures assignments in each position $i = 1, \ldots, P$. Note that procedures p=1,..., 6 are colonoscopy and $p = 7, \ldots, 10$ are combined upper endoscopy and colonoscopy procedures.

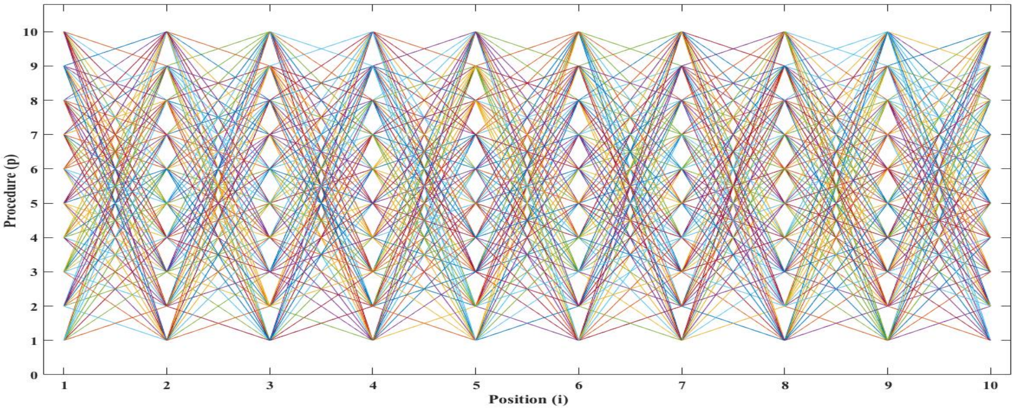


Figure 4.7: 1000 random sequences are presented

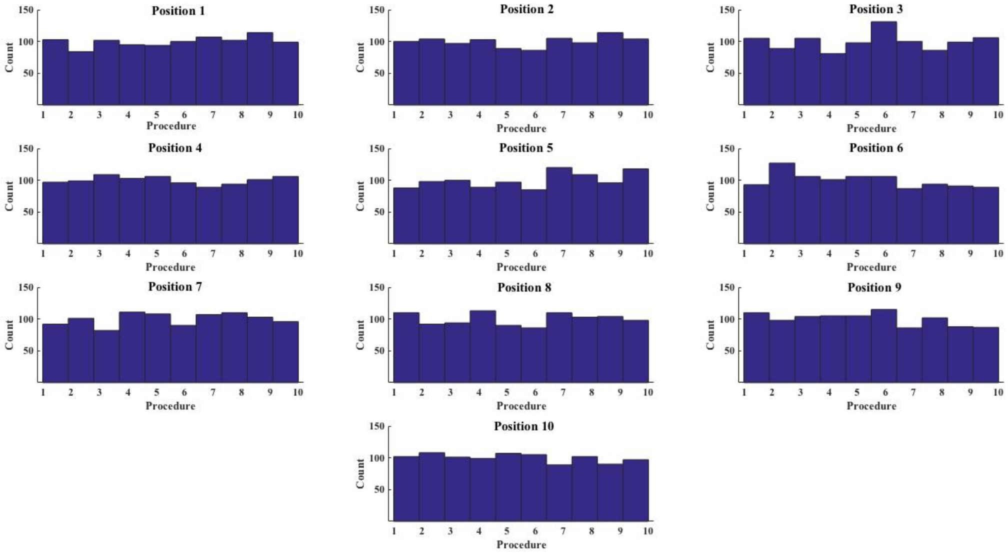
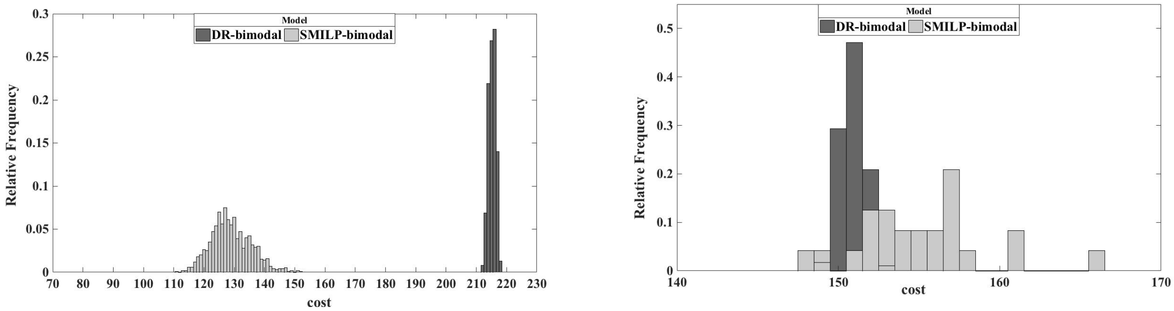


Figure 4.8: Histogram of procedures assignments in each position

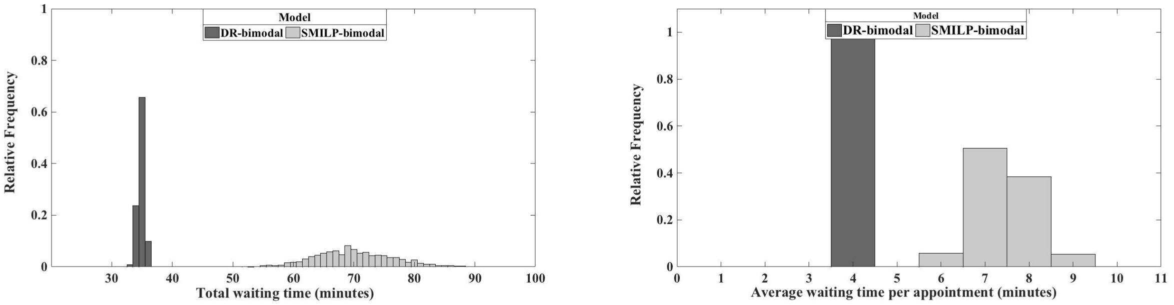
4.7.11 Cost of Revising The Optimal Sequence with Random Arrivals



(a) Optimal (in-sample) costs

(b) Out-of-sample simulation costs

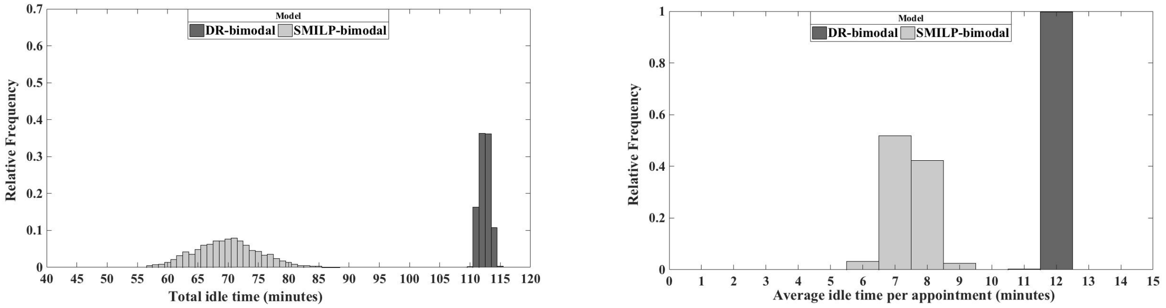
Figure 4.9: Relative frequency histograms of the optimal (in-sample) and out-of-sample simulation costs with random arrivals



(a) total waiting time

(b) waiting time per appointment

Figure 4.10: Relative frequency histograms of the out-of-sample waiting time with random arrivals.



(a) total provider idle time

(b) provider idle time per appointment

Figure 4.11: Relative frequency histograms of the out-of-sample idle time with random arrivals.

4.7.12 DR and SMILP Solution Time

In this section, we increase the size of the heterogeneous instances of the problem from P =10 to P=16, 20, and 30 procedures and compare the CPU time required for solving them using the DR-bimodal and SMILP-bimodal. In each instance, we fix the number of colonoscopy procedures (C) to $70\%P$, and the number of the combined upper endoscopy and colonoscopy (UC) to $25\%P$. For each of the resulting procedures mix, we generate five independent sets of the required random parameters for the DR and SMILP models as described in Section 4.5.1. We solve the DR-bimodal and SMILP-bimodal with each parameter sets of each procedure mix (instance). Table 4.9 presents the average CPU time (in second) (or the relative MIP= $\frac{\text{UB}-\text{LB}}{\text{UB}} \times 100\%$, where UB is the best upper bound and LB is the linear programming relaxation-based lower bound obtained at termination after 2 hours time limit) across the five solves of each instance using the DR-bimodal and SMILP-bimodal.

Table 4.9: Average CPU time (in second) of solving the DR-bimodal and SMILP-bimodal with various heterogeneous instances.

Instance	# of Procedures	Procedures Mix	DR-bimodal	SMILP-bimodal
1	10	(6C, 4UC)	1	123
2	16	(12C, 5UC)	7	600
3	20	(15C, 5UC)	80	9%
4	30	(23C, 7UC)	3987	40%

4.7.13 Out-of-Sampel Performance

Table 4.10: Out-of-sample performance of the optimal schedules given by DR and SLP models for the homogeneous instance (10 C) under perfect distributional information.

		Punctual Arrivals					
		$c^w = c^g = c^o$			$c^w = 1, c^g = 5, c^o = 7.5$		
Metric	Model	WT	OT	IT	WT	OT	IT
Mean	SLP-bimodal	7	6	5	17	2	1
	DR-bimodal	5	17	8	8	7	5
	SLP-plain	6	3	5	6	3	5
	DR-plain	9	4	4	18	2	2
Median	SLP-bimodal	0	0	0	8	0	0
	DR-bimodal	0	10	6	0	0	0
	SLP-plain	0	0	1	0	0	1
	DR-plain	0	0	0	8	0	0
75%-quantile	SLP-bimodal	8	0	9	26	0	0
	DR-bimodal	2	27	14	9	6	10
	SLP-plain	7	0	9	6	0	10
	DR-plain	11	0	8	28	0	0
95%-quantile	SLP-bimodal	37	38	17	62	13	9
	DR-bimodal	28	60	21	40	43	18
	SLP-plain	34	23	17	32	22	18
	DR-plain	42	29	16	67	16	10
		Random Arrivals					
		$c^w = c^g = c^o$			$\boldsymbol{c^w = 1, c^g = 5, c^o} = 7.5$		
Metric	Model	WT	OT	IT	WT	OT	IT
Mean	SLP-bimodal	7	9	6	18	3	2
	DR-bimodal	4	45	10	6	15	7
	SLP-plain	8	7	6	19	3	2
	DR-plain	6	11	7	15	4	3
Median	SLP-bimodal	0	0	3	10	0	0
	DR-bimodal	0	40	10	0	6	5
	SLP-plain	0	0	2	11	0	0
	DR-plain	0	2	5	6	0	0
75%-quantile	SLP-bimodal	10	11	11	28	0	2
	DR-bimodal	0	54	17	6	22	13
	SLP-plain	11	4	11	29	0	2
	DR-plain	7	16	13	23	0	6
95%-quantile	SLP-bimodal	37	48	19	63	24	13
	DR-bimodal	23	82	26	34	60	21
	SLP-plain	39	41	19	65	22	14
	DR-plain	34	50	21	60	31	16

Table 4.11: Out-of-sample performance of optimal schedules given by DR and SMLP models for the heterogeneous instance (6C, 4UC) under perfect distributional information.

		Punctual Arrivals					
		$c^{\text{w}} = c^{\text{g}} = c^{\text{o}}$			$c^{\text{w}} = 1, c^{\text{g}} = 5, c^{\text{o}} = 7.5$		
Metric	Model	WT	OT	IT	WT	OT	IT
Mean	SMILP-bimodal	9	0	3	15	0	2
	DR-bimodal	4	1	8	10	0	4
	SMILP-plain	10	0	3	19	0	1
	DR-plain	7	0	5	15	0	2
Median	SMILP-bimodal	0	0	0	6	0	0
	DR-bimodal	0	0	6	0	0	0
	SMILP-plain	0	0	0	11	0	0
	DR-plain	0	0	0	6	0	0
75%-quantile	SMILP-bimodal	13	0	6	23	0	0
	DR-bimodal	1	0	15	14	0	7
	SMILP-plain	14	0	5	29	0	0
	DR-plain	8	0	9	24	0	1
95%-quantile	SMILP-bimodal	43	0	14	57	0	10
	DR-bimodal	26	2	22	49	0	10
	SMILP-plain	43	0	14	64	0	7
	DR-plain	35	0	17	58	0	10
		Random Arrivals					
		$c^{\text{w}} = c^{\text{g}} = c^{\text{o}}$			$c^{\text{w}} = 1, c^{\text{g}} = 5, c^{\text{o}} = 7.5$		
Metric	Model	WT	OT	IT	WT	OT	IT
Mean	SMILP-bimodal	6	1	7	17	0	3
	DR-bimodal	2	5	12	9	1	6
	SMILP-plain	6	1	7	15	0	3
	DR-plain	5	1	8	14	0	3
Median	SMILP-bimodal	0	0	5	9	0	0
	DR-bimodal	0	0	11	0	0	2
	SMILP-plain	0	0	4	7	0	0
	DR-plain	0	0	7	5	0	0
75%-quantile	SMILP-bimodal	6	0	12	26	0	3
	DR-bimodal	0	5	18	12	0	11
	SMILP-plain	7	0	12	24	0	3
	DR-plain	4	0	14	21	0	6
95%-quantile	SMILP-bimodal	32	0	21	62	0	14
	DR-bimodal	21	30	28	42	0	21
	SMILP-plain	32	0	21	57	0	14
	DR-plain	28	0	23	54	0	16

Table 4.12: Out-of-sample performance of optimal schedules given by DR and SLP model for the homogeneous instance (10C) under misspecified distributional information. Procedure durations and arrival time deviations were sampled from the UM-MPU data.

		Punctual Arrivals					
		$c^{w} = c^{g} = c^{o}$			$c^{w} = 1, c^{g} = 5, c^{o} = 7.5$		
Metric	Model	WT	OT	IT	WT	OT	IT
Mean	SLP-bimodal	6	2	5	15	1	1
	DR-bimodal	4	9	8	6	3	5
	SLP-plain	6	2	5	15	1	2
	DR-plain	7	2	5	6	2	6
Median	SLP-bimodal	0	0	1	7	0	0
	DR-bimodal	0	1	7	0	0	1
	SLP-plain	0	0	2	0	0	2
	DR-plain	0	0	0	0	0	6
75%-quantile	SLP-bimodal	7	0	9	24	0	0
	DR-bimodal	0	14	14	7	0	10
	SLP-plain	6	0	10	6	0	10
	DR-plain	8	0	9	23	0	1
95%-quantile	SLP-bimodal	33	18	17	58	0	9
	DR-bimodal	24	41	22	34	23	18
	SLP-plain	32	19	18	31	19	19
	DR-plain	36	17	17	59	2	11
		Random Arrivals					
		$c^{w} = c^{g} = c^{o}$			$c^{w} = 1, c^{g} = 5, c^{o} = 7.5$		
Metric	Model	WT	OT	IT	WT	OT	IT
Mean	SLP-bimodal	8	4	6	18	1	2
	DR-bimodal	4	35	11	6	7	7
	SLP-plain	8	4	6	18	1	3
	DR-plain	6	7	7	14	2	4
Median	SLP-bimodal	0	0	0	12	0	0
	DR-bimodal	0	32	9	0	0	3
	SLP-plain	0	0	0	11	0	0
	DR-plain	0	0	3	6	0	0
75%-quantile	SLP-bimodal	10	0	12	28	0	0
	DR-bimodal	0	42	18	7	9	14
	SLP-plain	11	0	12	27	0	1
	DR-plain	8	9	14	22	0	5
95%-quantile	SLP-bimodal	36	27	24	63	6	15
	DR-bimodal	23	67	32	32	39	26
	SLP-plain	37	28	24	61	8	16
	DR-plain	33	37	27	55	17	18

Table 4.13: Out-of-sample performance of optimal schedules given by DR and SMLP model for the heterogeneous instance (6C, 4UC) under misspecified distributional information. Procedure durations and arrival time deviations were sampled from the UM-MPU data.

		Punctual Arrivals					
		$c^{\mathrm{w}} = c^{\mathrm{g}} = c^{\mathrm{o}}$			$c^{\mathrm{w}} = 1, c^{\mathrm{g}} = 5, c^{\mathrm{o}} = 7.5$		
Metric	Model	WT	OT	IT	WT	OT	IT
Mean	SMILP-bimodal	10	0	3	18	0	1
	DR-bimodal	4	0	8	11	0	3
	SMILP-plain	7	0	5	18	0	1
	DR-plain	10	0	3	15	0	2
Median	SMILP-bimodal	0	0	0	9	0	0
	DR-bimodal	0	0	6	1	0	0
	SMILP-plain	0	0	0	0	0	0
	DR-plain	0	0	1	6	0	0
75%-quantile	SMILP-bimodal	14	0	5	27	0	0
	DR-bimodal	1	0	14	16	0	6
	SMILP-plain	14	0	5	28	0	0
	DR-plain	7	0	9	23	0	1
95%-quantile	SMILP-bimodal	45	0	14	65	0	8
	DR-bimodal	26	0	21	50	0	15
	SMILP-plain	45	0	14	65	0	8
	DR-plain	36	0	17	60	0	10
		Random Arrivals					
		$c^{\mathrm{w}} = c^{\mathrm{g}} = c^{\mathrm{o}}$			$c^{\mathrm{w}} = 1, c^{\mathrm{g}} = 5, c^{\mathrm{o}} = 7.5$		
Metric	Model	WT	OT	IT	WT	OT	IT
Mean	SMILP-bimodal	7	0	6	18	0	3
	DR-bimodal	3	4	11	10	1	5
	SMILP-plain	7	0	7	18	0	3
	DR-plain	7	0	7	15	0	4
Median	SMILP-bimodal	0	5	1	11	0	0
	DR-bimodal	0	0	9	0	0	0
	SMILP-plain	0	0	1	10	0	0
	DR-plain	0	0	4	7	0	0
75%-quantile	SMILP-bimodal	9	0	13	27	0	1
	DR-bimodal	0	2	8	14	0	11
	SMILP-plain	9	0	13	27	0	1
	DR-plain	7	0	15	23	0	4
95%-quantile	SMILP-bimodal	35	0	53	62	0	15
	DR-bimodal	23	0	32	45	0	24
	SMILP-plain	36	0	25	62	0	15
	DR-plain	31	0	27	58	0	17

Table 4.14: Out-of-sample performance of optimal schedules given by DR and SLP model for the homogeneous instance (10C) under misspecified distributional information. Procedure durations were sampled from normal and weibull distributions, and arrival time deviations from a uniform distribution. WT, OT, IT are the averages over Cost1 and Cost2.

		Punctual Arrivals					
		Normal			Weibull		
Metric	Model	WT	OT	IT	WT	OT	IT
Mean	SLP-bimodal	16	0	2	15	1	2
	DR-bimodal	6	2	5	6	2	6
	SLP-plain	18	2	2	17	1	2
	DR-plain	19	1	2	18	1	2
Median	SLP-bimodal	10	0	0	8	0	0
	DR-bimodal	0	0	0	0	0	0
	SLP-plain	11	0	0	9	0	0
	DR-plain	11	0	0	10	0	0
75%-quantile	SLP-bimodal	25	0	0	24	0	0
	DR-bimodal	8	0	10	8	0	11
	SLP-plain	30	0	8	28	0	0
	DR-plain	29	0	0	29	0	0
95%-quantile	SLP-bimodal	55	0	12	24	0	0
	DR-bimodal	28	12	22	8	0	11
	SLP-plain	30	17	20	60	7	14
	DR-plain	64	10	12	62	8	13

		Random Arrivals					
		Normal			Weibull		
Metric	Model	WT	OT	IT	WT	OT	IT
Mean	SLP-bimodal	18	1	2	16	1	3
	DR-bimodal	5	6	7	5	6	8
	SLP-plain	17	1	3	19	2	2
	DR-plain	13	1	4	15	3	3
Median	SLP-bimodal	13	0	0	9	0	0
	DR-bimodal	0	0	4	0	0	5
	SLP-plain	11	0	0	12	0	0
	DR-plain	6	0	0	7	0	0
75%-quantile	SLP-bimodal	28	0	2	26	0	3
	DR-bimodal	6	8	13	6	8	14
	SLP-plain	26	0	2	30	0	2
	DR-plain	21	0	6	24	0	5
95%-quantile	SLP-bimodal	58	1	14	57	0	14
	DR-bimodal	24	26	25	27	30	25
	SLP-plain	55	2	14	62	12	14
	DR-plain	49	7	17	55	21	16

Table 4.15: Out-of-sample performance of optimal schedules given by DR and SMLP model for the heterogeneous instance (6C, 4UC) under misspecified distributional information. Procedure durations were sampled from normal and weibull distributions, and arrival time deviations from a uniform distribution. WT, OT, IT are the averages over Cost1 and Cost2.

		Punctual Arrivals					
		Normal			Weibull		
Metric	Model	WT	OT	IT	WT	OT	IT
Mean	SMILP-bimodal	14	0	3	14	0	4
	DR-bimodal	7	0	5	8	0	6
	SMILP-plain	15	0	2	14	0	2
	DR-plain	11	0	3	11	0	3
Median	SMILP-bimodal	8	0	0	6	0	0
	DR-bimodal	2	0	2	2	0	3
	SMILP-plain	9	0	0	7	0	0
	DR-plain	5	0	0	4	0	0
75%-quantile	SMILP-bimodal	22	0	4	23	0	6
	DR-bimodal	11	0	9	12	0	10
	SMILP-plain	23	0	2	22	0	3
	DR-plain	17	0	4	16	0	5
95%-quantile	SMILP-bimodal	50	0	15	53	0	17
	DR-bimodal	34	0	23	37	0	22
	SMILP-plain	50	0	12	49	0	13
	DR-plain	42	0	16	42	0	16
		Random Arrivals					
		Normal			Weibull		
Metric	Model	WT	OT	IT	WT	OT	IT
Mean	SMILP-bimodal	12	0	4	11	0	5
	DR-bimodal	6	2	8	6	1	8
	SMILP-plain	12	0	4	11	0	5
	DR-plain	9	0	5	10	0	5
Median	SMILP-bimodal	6	0	1	5	0	2
	DR-bimodal	0	0	5	0	0	5
	SMILP-plain	4	0	2	6	0	2
	DR-plain	7	0	1	5	0	3
75%-quantile	SMILP-bimodal	19	0	7	17	0	8
	DR-bimodal	7	1	14	7	0	15
	SMILP-plain	19	0	6	17	0	8
	DR-plain	15	0	9	15	0	9
95%-quantile	SMILP-bimodal	43	0	19	44	0	19
	DR-bimodal	28	9	27	29	8	26
	SMILP-plain	41	0	19	41	0	19
	DR-plain	36	0	21	38	0	20

CHAPTER 5

Conclusions

5.1 Summary and Contributions

In this dissertation, we study three stochastic outpatient scheduling problems and address challenges associated with solving each of them.

In Chapter 2, we propose a new stochastic mixed-integer linear programe (SMILP) for solving the problem of finding a sequence of appointment times for a set of procedures for a single provider (where each procedure has a known type and a random duration that follows a known probability distribution), minimizing a weighted sum of waiting, idle time, and overtime. We provide theoretical and empirical comparisons to other SMILP models in the literature, demonstrating where a significant improvement in computational performance can be gained with our model. More broadly, Chapter 2 presents the first rigorous and computational analysis of models for single-server stochastic appointment sequencing and scheduling with stochastic service duration, which has applications both within and outside of healthcare operations.

One of the shortcomings of the scheduling models that we analyze in Chapter 2 is that they ignore the uncertainty of patient's arrival times and the possibility of rescheduling (i.e., resequencing or declining). Therefore, in Chapter 3, we study an adaptive stochastic outpatient appointment scheduling problem (SOASP) which incorporates the random patients arrival times and random service durations, and adaptive rescheduling. Finding an optimal solution to this problem requires solving a multi-stage stochastic mixed-integer program (MSMIP) with an initial schedule made in the first stage and rescheduling policy optimized in the subsequent stages.

By deriving two-stage approximations of SOASP–MSMIP and testing them in extensive numerical experiments, we show that the easily-implementable appointment order policy (AO), which requires that patients are served in the order of their scheduled appointments, is near-optimal in a wide range of SOASP parameter settings. We also identify parameter settings that result in sub-optimality of the AO policy. Accordingly, we propose an alternative policy based on neighbor-swapping that improve the solutions of such instances.

To the best of our knowledge, and according to the recent review of outpatient appointment systems by *Ahmadi-Javid et al.* (2017), Chapter 3 presents the first stochastic programming approach to SOASP that considers (1) patient heterogeneity, (2) optimizing both the initial appointment sequencing and scheduling decisions, and (3) the possibility of rescheduling.

Chapters 2–3 both assume that we know the probability distributions of uncertain parameters; this isn't always the case. In Chapter 4, therefore, we consider the challenges of outpatient scheduling under under ambiguous probability distributions in the context of colonoscopy scheduling.

The main challenge in colonoscopy scheduling is that procedure duration depends on the quality of pre-procedure bowel prep that the patient must undergo. Data from a large outpatient clinic (OPC) indicates that colonoscopy durations are bimodal, i.e., depending on the prep quality they can follow two different probability distributions, one for those with adequate prep and the other for those with inadequate prep. Furthermore, there is a wide range of possible distributions for modeling the variability in colonoscopy duration with adequate and inadequate prep. This bimodality and ambiguity in the distribution of colonoscopy duration prevent us from using the stochastic programming approaches in Chapters 2–3. We therefore define a distributionally robust outpatient colonoscopy scheduling (DROCS) problem that seeks an optimal appointment sequence and schedule to minimize the worst-case weighted expected sum of patient waiting, provider idling, and provider overtime, where the worst-case is taken over an ambiguity set characterized through the known mean and support of the prep quality and durations.

We derive an equivalent mixed-integer linear programming formulation to solve DROCS. Finally, we present a case study based on extensive numerical experiments in which we draw several managerial insights into colonoscopy scheduling. According to the recent review of outpatient appointment systems by *Ahmadi-Javid et al.* (2017), the work in Chapter 4 is the first to address the bimodal ambiguity of colonoscopy (service) durations. We further contribute with a new DR model that incorporates sequencing decisions and considers the ambiguity of two coexisting uncertainties of colonoscopy duration (as a function of uncertain prep quality) and arrival time deviation.

Collectively, this dissertation addresses four salient challenges to efficient outpatient appointment scheduling under uncertainty: random service duration (Chapter 2), random arrival time (Chapter 3), the possibility of rescheduling (Chapter 3), and bimodality and ambiguity of the distribution of service duration (Chapter 4). More broadly, this dissertation contributes to the literature on scheduling under uncertainty and stochastic optimization with guidelines and methods to develop tractable and implementable scheduling (and mixed-integer programming) models and approaches.

5.2 Future Research

We suggest four areas for future research. First, we would like to extend our approach to include additional sources of uncertainty such as patient no-show, provider arrival time, and setup times between appointments. Second, we are interested in studying trade-offs between scheduling metrics such as provider workload (i.e., number of scheduled appointments) and patient access delays (i.e., the length of time a patient has to wait from the appointment request until a scheduled appointment becomes available to them). Third, our models and approaches assume a fixed number of patients with known types. We seek to use the results of this research to develop templates and policies for scheduling patients dynamically as they randomly request future appointments, considering patient and provider preferences, patient priority in terms of their respective medical urgency, and appointment utilization.

Finally, each patient in the scheduling problems we considered requires a single resource (e.g., a physician), and so we focused on optimizing appointment start time decisions. Some OPCs provide more complex services for elective outpatients, requiring several constrained resources such as a physician, a particular procedure room, specialized equipment, and one or more nurses. Some of these resources often require random setup time and have limited and unpredictable availability on each day. As such, OPC managers must choose a suitable set of resources and an appointment day, in addition to the appointment start time for each patient. Thus, we would like to extend our models and approach to multi-resource, integrated stochastic appointment planning and scheduling problems in which we need to assign a mode (resources), a day in the planning horizon, and an exact start time for each customer type.

BIBLIOGRAPHY

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

Ahmed, S., A. Shapiro, and E. Shapiro (2002), The sample average approximation method for stochastic programs with integer recourse, *Submitted for publication*, pp. 1–24.

Alexopoulos, C., D. Goldsman, J. Fontanesi, D. Kopald, and J. R. Wilson (2008), Modeling patient arrivals in community clinics, *Omega*, *36*(1), 33–43.

Almadi, M. A., M. Sewitch, A. N. Barkun, M. Martel, and L. Joseph (2015), Adenoma detection rates decline with increasing procedural hours in an endoscopists workload, *Canadian Journal of Gastroenterology and Hepatology*, *29*(6), 304–308.

Anderson, J. C., and L. F. Butterly (2015), Colonoscopy: quality indicators, *Clinical and translational gastroenterology*, *6*(2), e77.

Antunes, C. H., M. J. Alves, and J. Clímaco (2016), *Multiobjective linear and integer programming*, Springer.

Artigues, C., O. Koné, P. Lopez, and M. Mongeau (2015), Mixed-integer linear programming formulations, in *Handbook on Project Management and Scheduling Vol. 1*, edited by C. Schwindt and J. Zimmermann, pp. 17–41, Springer.

Bechtold, M. L., F. Mir, S. R. Puli, and D. L. Nguyen (2016), Optimizing bowel preparation for colonoscopy: a guide to enhance quality of visualization, *Annals of Gastroenterology: Quarterly Publication of the Hellenic Society of Gastroenterology*, *29*(2), 137.

Begen, M. A., and M. Queyranne (2011), Appointment scheduling with discrete random durations, *Mathematics of Operations Research*, *36*(2), 240–257.

Begen, M. A., R. Levi, and M. Queyranne (2012), A sampling-based approach to appointment scheduling, *Operations research*, *60*(3), 675–681.

Ben-Tal, A., and A. Nemirovski (1998), Robust convex optimization, *Mathematics of operations research*, *23*(4), 769–805.

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

Bertsimas, D., and I. Popescu (2005), Optimal inequalities in probability theory: A convex optimization approach, *SIAM Journal on Optimization*, *15*(3), 780–804.

Bertsimas, D., X. V. Doan, K. Natarajan, and C.-P. Teo (2010), Models for minimax stochastic linear optimization problems with risk aversion, *Mathematics of Operations Research*, *35*(3), 580–602.

Birge, J. R., and F. Louveaux (2011), *Introduction to stochastic programming*, Springer Science & Business Media.

Birge, J. R., and F. V. Louveaux (1988), A multicut algorithm for two-stage stochastic linear programs, *European Journal of Operational Research*, *34*(3), 384–392.

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

Boyd, S., and L. Vandenberghe (2004), *Convex optimization*, Cambridge university press.

Brahimi, M., and D. Worthington (1991), Queueing models for out-patient appointment systemsa case study, *Journal of the Operational Research Society*, *42*(9), 733–746.

Catanzaro, D., L. Gouveia, and M. Labbé (2015), Improved integer linear programming formulations for the job sequencing and tool switching problem, *European Journal of Operational Research*, *244*(3), 766–777.

Cayirli, T., and E. Veral (2003), Outpatient scheduling in health care a review of literature, *Production and operations management*, *12*(4), 519–549.

Cayirli, T., and K. K. Yang (2014), A universal appointment rule with patient classification for service times, no-shows, and walk-ins, *Service Science*, *6*(4), 274–295.

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

Cayirli, T., E. Veral, and H. Rosen (2008), Assessment of patient classification in appointment system design, *Production and Operations Management*, *17*(3), 338–353.

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

Cheong, S., R. R. Bitmead, and J. Fontanesi (2013), Modeling scheduled patient punctuality in an infusion center, *Lecture Notes in Management Science*, *5*, 46–56.

Chokshi, R. V., C. E. Hovis, T. Hollander, D. S. Early, and J. S. Wang (2012), Prevalence of missed adenomas in patients with inadequate bowel preparation on screening colonoscopy, *Gastrointestinal endoscopy*, *75*(6), 1197–1203.

Cox, T. F., J. P. Birchall, and H. Wong (1985), Optimising the queuing system for an ear, nose and throat outpatient clinic, *Journal of Applied Statistics*, *12*(2), 113–126.

Creemers, S., P. Colen, and M. Lambrecht (2012), Evaluation of appointment scheduling rules: a multi-performance measures approach, available at SSRN: http://dx.doi.org/10.2139/ssrn.2086264.

Deceuninck, M., D. Fiems, and S. De Vuyst (2018), Outpatient scheduling with unpunctual patients and no-shows, *European Journal of Operational Research*, *265*(1), 195–207.

Delage, E., and Y. Ye (2010), Distributionally robust optimization under moment uncertainty with application to data-driven problems, *Operations research*, *58*(3), 595–612.

Denton, B., and D. Gupta (2003), A sequential bounding approach for optimal appointment scheduling, *IIE Transactions*, *35*(11), 1003–1016.

Denton, B., J. Viapiano, and A. Vogl (2007), Optimization of surgery sequencing and scheduling decisions under uncertainty, *Health Care Management Science*, *10*(1), 13–24.

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

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

Esfahani, P. M., and D. Kuhn (2018), Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations, *Mathematical Programming*, *171*(1-2), 115–166.

Fetter, R. B., and J. D. Thompson (1966), Patients' waiting time and doctors' idle time in the outpatient setting, *Health services research*, *1*(1), 66.

Forst, F. G. (1993), Stochastic sequencing on one machine with earliness and tardiness penalties, *Probability in the Engineering and Informational Sciences*, *7*(2), 291–300.

Fortz, B., O. Oliveira, and C. Requejo (2017), Compact mixed integer linear programming models to the minimum weighted tree reconstruction problem, *European Journal of Operational Research*, *256*(1), 242–251.

Froehlich, F., V. Wietlisbach, J.-J. Gonvers, B. Burnand, and J.-P. Vader (2005), Impact of colonic cleansing on quality and diagnostic yield of colonoscopy: the european panel of appropriateness of gastrointestinal endoscopy european multicenter study, *Gastrointestinal endoscopy*, *61*(3), 378–384.

Gabrel, V., C. Murat, and A. Thiele (2014), Recent advances in robust optimization: An overview, *European journal of operational research*, *235*(3), 471–483.

Garey, M. R., D. S. Johnson, and R. Sethi (1976), The complexity of flowshop and jobshop scheduling, *Mathematics of operations research*, *1*(2), 117–129.

Ge, D., G. Wan, Z. Wang, and J. Zhang (2013), A note on appointment scheduling with piecewise linear cost functions, *Mathematics of Operations Research*, *39*(4), 1244–1251.

Glowacka, K. J., J. H. May, R. M. Goffman, E. K. May, A. S. Milicevic, K. L. Rodriguez, Y. C. Tjader, D. L. Vargas, and L. G. Vargas (2017), On prioritizing on-time arrivals in an outpatient clinic, *IISE Transactions on Healthcare Systems Engineering*, *7*(2), 93–106.

Gneezy, U., S. Meier, and P. Rey-Biel (2011), When and why incentives (don't) work to modify behavior, *Journal of Economic Perspectives*, *25*(4), 191–210.

Gul, S., B. T. Denton, J. W. Fowler, and T. Huschka (2011), Bi-criteria scheduling of surgical services for an outpatient procedure center, *Production and Operations management*, *20*(3), 406–417.

Gupta, D. (2007), Surgical suites' operations management, *Production and Operations Management*, *16*(6), 689–700.

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

Gupta, D., and W.-Y. Wang (2012), Patient appointments in ambulatory care, in *Handbook of healthcare system scheduling*, pp. 65–104, Springer.

Hanasusanto, G. A., D. Kuhn, and W. Wiesemann (2016), A comment on computational complexity of stochastic programming problems, *Mathematical Programming*, *159*(1-2), 557–569.

Hartigan, J. A., P. M. Hartigan, et al. (1985), The dip test of unimodality, *The annals of Statistics*, *13*(1), 70–84.

Ho, C.-J., and H.-S. Lau (1992), Minimizing total cost in scheduling outpatient appointments, *Management Science*, *38*(12), 1750–1764.

Homem-de Mello, T., and G. Bayraksan (2014), Monte carlo sampling-based methods for stochastic optimization, *Surveys in Operations Research and Management Science*, *19*(1), 56–85.

Jansson, B. (1966), Choosing a good appointment system-a study of queues of the type (d, m, 1), *Operations Research*, *14*(2), 292–312.

Jiang, B., J. Tang, and C. Yan (2019), A stochastic programming model for outpatient appointment scheduling considering unpunctuality, *Omega*, *82*, 70–82.

Jiang, R., S. Shen, and Y. Zhang (2017), Integer programming approaches for appointment scheduling with random no-shows and service durations, *Operations Research*, *65*(6), 1638–1656.

Johnson, D. A., et al. (2014), Optimizing adequacy of bowel cleansing for colonoscopy: recommendations from the us multi-society task force on colorectal cancer, *The American journal of gastroenterology*, *109*(10), 1528.

Jünger, M., T. M. Liebling, D. Naddef, G. L. Nemhauser, W. R. Pulleyblank, G. Reinelt, G. Rinaldi, and L. A. Wolsey (2009), *50 years of integer programming 1958–2008: From the early years to the state-of-the-art*, Springer Science & Business Media.

Keha, A. B., K. Khowala, and J. W. Fowler (2009), Mixed integer programming formulations for single machine scheduling problems, *Computers & Industrial Engineering*, *56*(1), 357–367.

Kim, S., R. Pasupathy, and S. G. Henderson (2015), A guide to sample average approximation, in *Handbook of Simulation Optimization*, edited by M. C. Fu, pp. 207–243, Springer.

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

Klassen, K. J., and R. Yoogalingam (2014), Strategies for appointment policy design with patient unpunctuality, *Decision Sciences*, *45*(5), 881–911.

Kleywegt, A. J., A. Shapiro, and T. Homem-de Mello (2002), The sample average approximation method for stochastic discrete optimization, *SIAM Journal on Optimization*, *12*(2), 479–502.

Klotz, E., and A. M. Newman (2013), Practical guidelines for solving difficult linear programs, *Surveys in Operations Research and Management Science*, *18*(1–2), 1–17.

Kocas, C. (2015), An extension of osunas model to observable queues, *Journal of Mathematical Psychology*, *66*, 53–58.

Kong, Q., C.-Y. Lee, C.-P. Teo, and Z. Zheng (2013), Scheduling arrivals to a stochastic service delivery system using copositive cones, *Operations research*, *61*(3), 711–726.

Kong, Q., S. Li, N. Liu, C.-P. Teo, and Z. Yan (2015), Appointment scheduling under schedule-dependent patient no-show behavior.

Lawler, E. L., J. K. Lenstra, A. H. R. Kan, and D. B. Shmoys (1993), Sequencing and scheduling: Algorithms and complexity, *Handbooks in Operations Research and Management Science*, *4*, 445–522.

Lebwohl, B., T. C. Wang, and A. I. Neugut (2010), Socioeconomic and other predictors of colonoscopy preparation quality, *Digestive diseases and sciences*, *55*(7), 2014–2020.

Linderoth, J., A. Shapiro, and S. Wright (2006), The empirical behavior of sampling methods for stochastic programming, *Annals of Operations Research*, *142*(1), 215–241.

Lindley, D. V. (1952), The theory of queues with a single server, in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 48, pp. 277–289, Cambridge University Press.

Maechler, M. (2016), Package 'diptest'.

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

Mak, W.-K., D. P. Morton, and R. K. Wood (1999), Monte carlo bounding techniques for determining solution quality in stochastic programs, *Operations research letters*, *24*(1-2), 47–56.

Mancilla, C., and R. Storer (2012), A sample average approximation approach to stochastic appointment sequencing and scheduling, *IIE Transactions*, *44*(8), 655–670.

Marler, R. T., and J. S. Arora (2004), Survey of multi-objective optimization methods for engineering, *Structural and multidisciplinary optimization*, *26*(6), 369–395.

Mercer, A. (1960), A queueing problem in which the arrival times of the customers are scheduled, *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 108–113.

Mittal, S., A. S. Schulz, and S. Stiller (2014), Robust appointment scheduling, in *LIPIcs-Leibniz International Proceedings in Informatics*, vol. 28, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Molina-Pariente, J. M., E. W. Hans, and J. M. Framinan (2016), A stochastic approach for solving the operating room scheduling problem, *Flexible services and manufacturing journal*, pp. 1–28.

Morales-España, G., C. M. Correa-Posada, and A. Ramos (2016), Tight and compact mip formulation of configuration-based combined-cycle units, *IEEE Transactions on Power Systems*, *31*(2), 1350–1359.

Okotie, O. T., N. Patel, and C. M. Gonzalez (2008), The effect of patient arrival time on overall wait time and utilization of physician and examination room resources in the outpatient urology clinic, *Advances in urology*, *2008*.

Ostrowski, J., J. Linderoth, F. Rossi, and S. Smriglio (2011), Orbital branching, *Mathematical Programming*, *126*(1), 147–178.

Osuna, E. E. (1985), The psychological cost of waiting, *Journal of Mathematical Psychology*, *29*(1), 82–105.

Pinedo, M. L. (2016), *Scheduling: theory, algorithms, and systems*, Springer.

Pinto, J. M., and I. E. Grossmann (1998), Assignment and sequencing models for thescheduling of process systems, *Annals of Operations Research*, *81*, 433–466.

Pochet, Y., and L. A. Wolsey (2006), *Production planning by mixed integer programming*, Springer Science & Business Media.

Rex, D. K., T. F. Imperiale, D. R. Latinovich, and L. L. Bratcher (2002), Impact of bowel preparation on efficiency and cost of colonoscopy, *The American journal of gastroenterology*, *97*(7), 1696–1700.

Rex, D. K., et al. (2006), Quality indicators for colonoscopy, *The American journal of gastroenterology*, *101*(4), 873.

Riise, A., C. Mannino, and L. Lamorgese (2016), Recursive logic-based benders decomposition for multi-mode outpatient scheduling, *European Journal of Operational Research*, *255*(3), 719–728.

Robinson, L. W., and R. R. Chen (2003), Scheduling doctors' appointments: optimal and empirically-based heuristic policies, *Iie Transactions*, *35*(3), 295–307.

Rockafellar, R. T., and R. J.-B. Wets (1991), Scenarios and policy aggregation in optimization under uncertainty, *Mathematics of operations research*, *16*(1), 119–147.

Rohleder, T. R., and K. J. Klassen (2000), Using client-variance information to improve dynamic appointment scheduling performance, *Omega*, *28*(3), 293–302.

Sabria, F., and C. F. Daganzo (1989), Approximate expressions for queueing systems with scheduled arrivals and established service order, *Transportation Science*, *23*(3), 159–165.

Salzarulo, P. A., S. Mahar, and S. Modi (2016), Beyond patient classification: Using individual patient characteristics in appointment scheduling, *Production and Operations Management*, *25*(6), 1056–1072.

Samorani, M., and S. Ganguly (2016), Optimal sequencing of unpunctual patients in high-service-level clinics, *Production and Operations Management*, *25*(2), 330–346.

Scarf, H. (1958), A min max solution of an inventory problem, stud ies in the mathematical theory of inventory and production (chap ter 12).

Shang, C., and F. You (2018), Distributionally robust optimization for planning and scheduling under uncertainty, *Computers & Chemical Engineering*, *110*, 53–68.

Shapiro, A. (2003), Monte carlo sampling approach to stochastic programming, in *ESAIM: Proceedings*, vol. 13, pp. 65–73, EDP Sciences.

Shapiro, A., D. Dentcheva, and A. Ruszczyński (2009), *Lectures on stochastic programming: modeling and theory*, SIAM.

Shehadeh, K. S., A. E. Cohn, and M. A. Epelman (2019), Analysis of models for the stochastic outpatient procedure scheduling problem, *European Journal of Operational Research*.

Sheppard, M. (2012), `allfitdist` function, GitHub repository.

Singh, S., M. Dhawan, M. Chowdhry, M. Babich, and E. Aoun (2016), Differences between morning and afternoon colonoscopies for adenoma detection in female and male patients, *Annals of Gastroenterology: Quarterly Publication of the Hellenic Society of Gastroenterology*, *29*(4), 497.

Smith, J. E., and R. L. Winkler (2006), The optimizers curse: Skepticism and postdecision surprise in decision analysis, *Management Science*, *52*(3), 311–322.

Soriano, A. (1966), Comparison of two scheduling systems, *Operations Research*, *14*(3), 388–397.

Tai, G., and P. Williams (2012), Optimization of scheduling patient appointments in clinics using a novel modelling technique of patient arrival, *Computer methods and programs in biomedicine*, *108*(2), 467–476.

T'kindt, V., and J.-C. Billaut (2006), *Multicriteria scheduling: theory, models and algorithms*, Springer Science & Business Media.

Vanden Bosch, P. M., and D. C. Dietz (2001), Scheduling and sequencing arrivals to an appointment system, *Journal of Service Research*, *4*(1), 15–25.

Vissers, J., and J. Wijngaard (1979), The outpatient appointment system: Design of a simulation study, *European Journal of Operational Research*, *3*(6), 459–463.

Wagner, H. M. (1959), An integer linear-programming model for machine scheduling, *Naval Research Logistics (NRL)*, *6*(2), 131–140.

Weiss, E. N. (1990), Models for determining estimated start times and case orderings in hospital operating rooms, *IIE transactions*, *22*(2), 143–150.

Welch, J., and N. J. Bailey (1952), Appointment systems in hospital outpatient departments, *The Lancet*, *259*(6718), 1105–1108.

White, M. B., and M. Pike (1964), Appointment systems in out-patients' clinics and the effect of patients' unpunctuality, *Medical Care*, pp. 133–145.

Wiesemann, W., D. Kuhn, and M. Sim (2014), Distributionally robust convex optimization, *Operations Research*, *62*(6), 1358–1376.

Williams, K. A., C. G. Chambers, M. Dada, J. C. McLeod, and J. A. Ulatowski (2014), Patient punctuality and clinic performance: observations from an academic-based private practice pain centre: a prospective quality improvement study, *BMJ open*, *4*(5), e004,679.

Zacharias, C., and T. Yunes (2018), Multimodularity in the stochastic appointment scheduling problem with discrete arrival epochs, *Tech. rep.*, Working paper. School of Business Administration, University of Miami, Coral .

Zauber, A. G., et al. (2012), Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths, *New England Journal of Medicine*, *366*(8), 687–696.

Zhang, Y., S. Shen, and S. A. Erdogan (2017), Distributionally robust appointment scheduling with moment-based ambiguity set, *Operations Research Letters*, *45*(2), 139–144.

Zhu, H., Y. Chen, E. Leung, and X. Liu (2017), Outpatient appointment scheduling with unpunctual patients, *International Journal of Production Research*, pp. 1–21.