TECHNISCHE UNIVERSITÄT MÜNCHEN

TUM School of Management

Appointment Scheduling and Routing in Healthcare

Introducing Robust Service Levels

Carolin Isabel Bauerhenne

Vollständiger Abdruck der von der TUM School of Management der Technischen Universität München zur Erlangung einer

Doktorin der Wirtschafts- und Sozialwissenschaften
(Dr. rer. pol.)

genehmigten Dissertation.

Vorsitz:		Prof. Dr. Martin Bichler
Prüfende der Dissertation:	1.	Prof. Dr. Rainer Kolisch
	2.	Prof. Dr. Andreas S. Schulz
	3.	Prof. Dr. Christina Büsing

Die Dissertation wurde am 26.06.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Management am 15.08.2024 angenommen.

Appointment Scheduling and Routing in Healthcare

Introducing Robust Service Levels

Carolin Isabel Bauerhenne

Acknowledgments

First and foremost, I express my sincere gratitude to my advisors Prof. Dr. Rainer Kolisch and Prof. Dr. Andreas S. Schulz for their invaluable inspiration, guidance, encouragement, and unwavering support. Additionally, I would like to thank Prof. Dr. Jonathan Bard for his fruitful inspiration, guidance, and feedback, as well as for the great opportunity to initiate our collaboration with a research stay in Austin. Special appreciation goes to Prof. Dr. Christina Büsing for her role as an examiner and Prof. Dr. Martin Bichler for his role as chairman. My gratitude extends to Bernhard Renger, Dr. Knud Nairz, and Dr. Wolfram Bosbach for providing valuable real-world data and insights, and to Christine Steinberger, Anne Diehl, and Dr. Isabel Koch for their great administrative support.

My sincere thanks go to the Deutsche Forschungsgemeinschaft (DFG) - GRK 2201 for funding my research. Moreover, I am very grateful to the Bavarian Research Institute for Digital Transformation and the Simons Laufer Mathematical Sciences Institute for providing scholarships that supported my participation in various greatly inspiring workshops and conferences.

My heartfelt gratitude goes to all of my colleagues at the Chair of Operations Management, the Chair of Operations Research, and the research training group "Advanced Optimization in a Networked Economy" at the Technical University of Munich for fruitful conversations, assistance, and camaraderie. Furthermore, I am indebted to those who ignited my passion for research, particularly my mentor PD Dr. René Brandenberg, along with Prof. Dr. Stefan Weltge, Prof. Dr. Peter Gritzmann, and all of my Discrete Optimization colleagues.

Finally, I express profound gratitude to my parents, Margit and Karl-Peter, whose unwavering support has been a constant source of strength and joy. I would also like to thank my siblings, Julia, Daniel, and Stefanie, as well as their partners and my beloved nieces. My heartfelt gratitude further goes to all my friends, with special acknowledgment to Juliana, Marie, Tassilo, Isabelle, and Laura, for their invaluable acts of support and encouragement. Special appreciation also goes to Lilian, Baturhan, Sude, and Eleni for their wonderful companionship.

Thank you very much.

Abstract

Amidst the numerous operational challenges in healthcare delivery, appointment scheduling is responsible for managing patient waiting times, idle times, and overtime. Balancing these factors under uncertainty poses significant challenges in both research and practice. This dissertation introduces *robust service levels* that guarantee acceptable waiting times for each patient, even in a worst-case scenario. Under these constraints, worst-case costs are minimized and thereby capacity utilization is maximized.

The integration of robust service levels is applied to three operational problems: a general appointment scheduling framework, predictive overbooking, and home healthcare routing and scheduling. Robustness is ensured with respect to polyhedral uncertainty sets for service and travel time uncertainty, as well as with respect to novel uncertainty sets for false positive and false negative no-show predictions.

Based on decomposition approaches, a mixed-integer linear program is derived for each of these $\mathcal{NP}$-hard problems. The home healthcare routing and scheduling problem is then addressed using a nested branch-and-price procedure. Easy-to-implement optimal scheduling rules are presented for the other two problems.

Computational studies based on real-world data demonstrate the effectiveness of robust waiting time guarantees in ensuring more equitable and reliable patient waiting times while enabling precise cost management. The cost implications of these guarantees naturally vary depending on the chosen service levels. Significantly, compared to a traditional weighted sum approach, adopting robust service levels may even lead to a simultaneous reduction in costs.

Overall, this dissertation provides a comprehensive framework and solution methods for implementing robust service levels in appointment scheduling and routing. It demonstrates the benefits of guaranteeing acceptable worst-case waiting times on an individual patient basis to advance equity in appointment scheduling and routing, while minimizing costs.

Contents

List of Tables

List of Figures

1. Introduction

On the path toward enhancing healthcare delivery, operations management is responsible for optimally allocating scarce resources - such as facility capacities, equipment, pharmaceuticals, and caregivers. Thereby, operational decisions exert substantial influence on overall costs, healthcare access, and patient outcomes. Akin to patient-level clinical factors (Kc et al. 2020), these decisions carry profound ethical and economic implications, presenting both risks and opportunities.

Since before 1980, numerous scholarly publications have contributed to operations research in healthcare (Fries 1976, Krischer 1980). The allocation of spatial resources, including ambulance stations, operating rooms, and hospital beds, as well as of medical supplies such as vaccines, donor organs, blood collection, and pharmaceuticals, along with human resources, medical decision-making, and time has been examined across diverse healthcare settings. Methodological approaches span from empirical investigations and behavioral experiments to machine learning and optimization. Excellent and comprehensive reviews of established decision-making tools and managerial insights are, amongst others, provided by Keskinocak and Savva (2020), Kuo et al. (2020), and Dai and Tayur (2020).

Despite remarkable progress, it is evident that numerous operational challenges persist in healthcare delivery. Meanwhile, evolving circumstances require corresponding shifts in approaching these challenges. For instance, with the increasing availability of data, there is a growing need to adapt algorithms to be more data-driven (Kc et al. 2020). At the same time, most practitioners continue to demand that solutions are easy to implement (Keskinocak and Savva 2020). Recent emphasis has been placed on elevating care quality and fostering equity (Porter et al. 2009, Ayer et al. 2014, Keskinocak and Savva 2020).

1.1. Appointment Scheduling and Routing in Healthcare

Amid the myriad of challenges facing healthcare delivery, time-related issues persist globally and in various healthcare settings such as primary care, specialty care, and home healthcare. These issues include unacceptable patient waiting times for treatment (OECD 2020), inequitable capacity utilization (Morris et al. 2005, Hirello et al. 2022, Lampert 2021), excessive caregiver overtime (Marburger Bund 2013, Statistics Canada 2023), and unacceptable patient waiting times on the day of service (Kassenärztliche Bundesvereinigung 2021, Hill and Joonas 2005). Time management in healthcare is not only economically relevant for profitability and stakeholder satisfaction. Most importantly, time-related service quality and timely access also impact patient health outcomes (Aiken et al. 2002, Bae and Fabry 2014, Hermida et al. 2016).

At an operational level, appointment scheduling is the key tool for time management. Its primary responsibility includes managing patient waiting time, idle time, and overtime. The main decision revolves around determining appointment times for a group of patients seeking services within a specified time frame. Other factors to consider include patient acceptance and sequencing, amongst others. Notably, in home healthcare, appointment scheduling is intricately linked to routing decisions.

Starting with the pioneering works of Bailey (1952) and Lindley (1952), appointment scheduling has attracted the interest of many academicians and practitioners. The research landscape offers a wide array of valuable decision-support tools and managerial insights, as excellently reviewed by Cayirli and Veral (2003), Gupta and Denton (2008), Ahmadi-Javid et al. (2017), and Marynissen and Demeulemeester (2019). While optimal solutions have been demonstrated to be context-sensitive, three main sources of complexity persistently contribute to the difficulty of appointment scheduling and routing:

Uncertainty. Waiting times, idle times, and overtime are originally caused by the inherent uncertainty in factors such as service times and no-shows. Even in static settings, where the available information regarding the uncertainty does not change over time, there are three main frameworks for modeling different available

information: stochastic programming, distributionally robust optimization, and robust optimization.

Stochastic programming, historically the first approach, assumes knowledge of the probability distribution of all uncertain parameters and typically optimizes the expected value of the objective function, as outlined in Birge and Louveaux (2011) and Shapiro et al. (2021). Single-stage and two-stage stochastic programming are commonly used for appointment scheduling problems, see Cayirli and Veral (2003), Gupta and Denton (2008), and Ahmadi-Javid et al. (2017).

In various settings, the probability distribution of uncertain parameters may not be precisely known. In such cases, distributionally robust optimization approaches may be employed, optimizing the (typically) expected performance over the worst case across a set of plausible distributions, as thoroughly reviewed in Rahimian and Mehrotra (2022). These ambiguity sets can be constructed based on various criteria such as moments, likelihood, or Wasserstein distance. While the resulting models can be challenging in terms of tractability, they have been successfully applied to appointment scheduling, such as in Kong et al. (2013), Mak et al. (2015), Zhang et al. (2017), Jiang et al. (2017, 2019), and Kong et al. (2020).

However, relying solely on expected performance or similar probabilistic functions may not accurately model the individual patient experience. Instead, robust optimization offers to prevent detrimental singular outcomes by optimizing the worst-case performance for any realization within an uncertainty set based only on the support of the probability distribution of the uncertain factors. It has significant contributions from El Ghaoui and Lebret (1997), El Ghaoui et al. (1998), Ben-Tal and Nemirovski (1998), Bertsimas and Sim (2004), Ben-Tal and Nemirovski (2000), and Ben-Tal et al. (2009), as extensively reviewed by Lu and Shen (2021). Notably, the level of conservatism with this approach can be controlled through uncertainty budgets (Bertsimas and Sim 2004). While robust optimization has found limited application in appointment scheduling and routing, as reviewed in Sections 2.2.1 and 4.2.2, it may offer intuitive and tractable solutions to prepare healthcare time management against unfortunate scenarios.

Multiple Criteria. Among the wide range of available approaches to multi-objective optimization, as excellently outlined by Deb et al. (2016), the weighted

sum method has stood out as predominant in the appointment scheduling literature. Typically, this method has been used to minimize the expected value of a weighted sum of waiting times, overtime, and idle times. While this approach has offered valuable insights, it has also long been recognized that the aggregated consideration of waiting times and overtime may result in unfair outcomes (Cayirli and Veral 2003, Gupta and Denton 2008, Ahmadi-Javid et al. 2017, Keskinocak and Savva 2020). Section 2.2.2 provides a review of the few approaches focusing on equity and fairness. Given that waiting times affect individual patients, more nuanced and equitable approaches are necessary (Cayirli and Veral 2003, Gupta and Denton 2008, Keskinocak and Savva 2020).

Solvability. The complex nature of appointment scheduling and routing, a multi-criteria problem under uncertainty, often classified as $\mathcal{NP}$-hard (Bosch 1997, Kong et al. 2016, Schulz and Udwani 2019), requires advanced solution methods. Numerical methods, such as sample average approximation, simulation-based optimization, approximate dynamic programming algorithms, and metaheuristics, have been predominantly utilized, providing solutions iteratively. Exact solutions, if achieved, typically involve decomposition approaches or specialized software packages. In contrast, closed-form optimal solutions are rare (Cayirli and Veral 2003, Gupta and Denton 2008, Ahmadi-Javid et al. 2017). Ahmadi-Javid et al. (2017) emphasize the need to further explore efficient and accurate solution methods across various healthcare settings. Meanwhile, solutions should be easily implementable for practitioners (Cayirli and Veral 2003, Keskinocak and Savva 2020).

Hence, as the quest to enhance appointment scheduling and routing solutions continues, there is, amongst others, a need for easy-to-implement and robust solutions to improve service quality and promote equity.

1.2. A New Perspective: Robust Service Levels

Appointment scheduling is responsible for navigating the conflicting criteria of capacity utilization, overtime, and waiting times. Capacity utilization contributes to timely access to healthcare and is financially incentivized. Meanwhile, waiting times and overtime also provably contribute to health outcomes, as well as patient and caregiver satisfaction (Aiken et al. 2002, Bae and Fabry 2014, Hermida et al. 2016). Excessive amounts of any factor are harmful. Conversely, while minimizing these factors is desirable, it may not always be necessary (Huang 1994, Hill and Joonas 2005, Camacho et al. 2006). In this context, this dissertation aims to optimize appointment scheduling by reconciling these conflicting criteria in such a way that no criterion is prioritized at the expense of detrimental consequences from other criteria. Maximizing capacity utilization should not cause excessive waiting times or overtime, and minimizing waiting times should not unduly affect capacity utilization.

Hence, this dissertation introduces a novel concept to appointment scheduling and routing: *Robust service levels*. Splitting the traditional weighted sum objective function, waiting time guarantees are implemented as hard constraints to ensure an acceptable service level for each patient. These levels can be customized based on empirical thresholds for optimal health outcomes and patient satisfaction. As patients individually evaluate performance, typically based on a single or few experiences rather than aggregate statistics, a robust optimization approach is adopted to ensure these service levels for each patient not only on average but in worst-case scenarios. Within this patient-centered framework, cost is minimized, thereby maximizing capacity utilization.

The computational findings in this dissertation, based on real-world data, demonstrate that robust service levels can ensure acceptable waiting times for nearly all patients. Furthermore, waiting times are distributed more equitably compared to traditional methods. While robust service levels increase costs in some settings, they concurrently reduce costs in others.

Notably, robust service levels can even facilitate theoretical analysis compared to traditional methods, as evidenced by the optimal scheduling and sequencing rules outlined in Section 2.4. They also enable a seamless integration of prescrip-

tive analytic findings, as demonstrated in Chapter 3.

Overall, in pursuit of advancing equity in patient care, the proposed concept for introducing robust service levels offers intuitive and easy-to-implement solutions while concurrently optimizing capacity utilization. Its applicability extends beyond appointment scheduling to other problem domains and application areas.

1.3. Research Problems

In this dissertation, robust service levels are applied to the following problems:

Intraday Appointment Scheduling. The first problem under investigation is a general framework for offline appointment scheduling of preselected patients with a single service provider, aiming to decide an optimal patient sequence and appointment times. A key challenge involves addressing uncertainty in service times, defined by box uncertainty sets. The distinctive feature introduced in this dissertation is the incorporation of robust waiting time guarantees for all patients. While ensuring these guarantees as hard constraints, the objective is to minimize costs related to idle time and overtime. The study analytically derives optimal and easy-to-implement sequencing and scheduling methods, subsequently evaluated using CT data from a radiology department at a major university hospital in Munich, Germany.

Predictive Overbooking. The second problem extends the appointment scheduling problem with waiting time guarantees to account for no-shows. Leveraging binary predictions regarding individual no-show behavior, this dissertation introduces uncertainty budgets for false positives and false negatives. Waiting time guarantees are established for the resulting overbooking problem, ensuring adherence even under worst-case scenarios. Using decomposition, total unimodularity, and LP duality, a mixed-integer linear program is formulated. Optimal solutions are assessed via simulation based on real-world CT data from a major university hospital in Munich, Germany.

Home Healthcare Routing and Scheduling. The third problem involves a home healthcare agency simultaneously selecting new patients, assigning them

to caregivers, and creating robust weekly routes and schedules over a multi-day planning horizon with uncertain travel and service times. The proposed model incorporates various optimization criteria, including patient time windows, number of required visits, caregiver availability, compatibility, continuity of care, and profit; with the distinctive feature that all patient-related criteria are modeled as hard constraints. This entails designing robust schedules to ensure timely care and prevent overtime, while controlling the level of conservatism by uncertainty budgets. To address this problem, a nested branch-and-bound procedure is proposed and evaluated using real-world data from a US-based home healthcare agency, leading to optimal decision patterns derived from an extensive case study.

1.4. Outline

The structure of this dissertation is as follows:

Chapter 2 addresses robust appointment scheduling. Section 2.1 motivates the introduction of waiting time guarantees based on empirical thresholds for acceptable waiting times. Section 2.2 reviews related optimization problems in the appointment scheduling literature. In Section 2.3, the robust appointment scheduling problem with waiting time guarantees is formally introduced, proving its $\mathcal{NP}$-hardness and providing a mixed-integer linear program formulation. Section 2.4 presents polynomial-time optimal scheduling and sequencing rules for relevant special cases. Subsequently, in Section 2.5, a computational study is conducted using real-world data to evaluate the cost implications of implementing maximum waiting time guarantees.

Chapter 3 extends the robust appointment scheduling problem with waiting time guarantees to consider no-shows, resulting in an overbooking problem. Section 3.1 motivates the consideration of individual no-show predictions for overbooking. Section 3.2 reviews related problems in the literature. Section 3.3 introduces the Robust Predictive Overbooking problem with Waiting Time Guarantees and derives a mixed-integer linear program formulation by exploiting the problem structure. Section 3.4 presents a simulation based on real-world data, demonstrating that predictive overbooking can significantly increase capacity utilization without compromising waiting time guarantees.

Chapter 4 addresses the problem of home healthcare routing and scheduling.

Section 4.1 underlines the need for a robust approach to all health-related criteria. Section 4.2 reviews related optimization problems in the home healthcare routing and scheduling literature. In Section 4.3, a robust mixed-integer linear program for weekly home healthcare routing and scheduling is introduced, followed by the development of a nested branch-and-price algorithm in Section 4.4 to efficiently solve the problem for realistic instances. Finally, in Section 4.5, a computational study is conducted using real-world data, providing insights into profit sensitivity and optimal decision patterns for different levels of robustness.

2. Robust Appointment Scheduling with Waiting Time Guarantees

Co-authors: Rainer Kolisch and Andreas S. Schulz
Working paper (*Resubmitted* at *Manufacturing & Service Operations Management* at the time of submission of this dissertation).

Abstract. Appointment scheduling problems under uncertainty encounter a fundamental trade-off between cost minimization and patient waiting times. Most existing studies address this trade-off using a weighted sum approach, which puts little emphasis on individual waiting times and, thus, patient satisfaction. In contrast, we study how to minimize total cost while providing waiting time guarantees to all patients. Specifically, we introduce the Robust Appointment Scheduling Problem with Waiting Time Guarantees, taking into account uncertainty in service times represented by box uncertainty sets. We show that the problem is $\mathcal{NP}$-hard in general and introduce a mixed-integer linear program that can be solved in reasonable computation time. For special cases, we prove that polynomial-time variants of the well-known Smallest-Variance-First sequencing rule and the Bailey-Welch scheduling rule are optimal. Furthermore, a case study with data from the radiology department of a large university hospital demonstrates that the approach not only guarantees acceptable waiting times but, compared to existing robust approaches, may simultaneously reduce costs incurred by idle time and overtime. Our findings suggests that limiting instead of minimizing customer waiting times is a win-win solution in the trade-off between patient satisfaction and cost minimization. Additionally, it provides an easy-to-implement and customizable appointment scheduling framework with waiting time guarantees.

2.1. Introduction

Excessive waiting times negatively affect patient satisfaction, perceived service quality, and loyalty (Huang 1994, Taylor and Baker 1994). A patient's tolerance for waiting, which separates acceptable from excessive waiting times, has been shown to range from 37 minutes for a majority of patients in an outpatient clinic to 11 seconds for most users to load a website (Bouch et al. 2000, Moschis et al. 2003, Hill and Joonas 2005). The tolerance depends amongst others on the perceived and expected waiting time and can be influenced by many effective interventions, e.g., Kumar et al. (1997), Davis and Heineke (1998), Bielen and Demoulin (2007). Meanwhile, there is a lack of appointment scheduling approaches to control actual waiting times.

Appointment scheduling is a multi-criteria problem that typically involves uncertainty in service times. In particular, appointment scheduling problems face a fundamental trade-off between waiting times and capacity utilization: Shorter waiting times come at the cost of longer idle times, and vice versa. In the literature, this trade-off has been modeled predominantly by minimizing a weighted sum of waiting times, idle times, and overtime (Ahmadi-Javid et al. 2017). However, it has long been known that such weighted sum approaches can lead to unbalanced and excessive waiting times for some patients (Cayirli and Veral 2003), and in combination with individual no-show predictions even to racial disparity (Samorani et al. 2021).

This chapter proposes a new approach to appointment scheduling that considers waiting (in)tolerance. In particular, we study the following robust appointment scheduling and sequencing problem. For a set of preselected patients, we are given the shortest and longest possible service time of each patient. We limit each patient's waiting time, even in a worst-case scenario, to a predetermined tolerance for waiting (so-called maximum waiting time guarantee). Then we minimize a weighted sum of idle times and overtime (so-called total cost). The main contributions of this work can be summarized as follows.

Maximum Waiting Time Guarantees. To the best of our knowledge, this work is the first to introduce maximum waiting time guarantees for intraday appointment scheduling. The guarantees ensure that each patient's waiting time

stays below a given individual threshold, even in a worst-case scenario. Hence, they ensure that all patients experience acceptable waiting times and prevent patient dissatisfaction due to excessive waiting times. In addition, communicating these guarantees can increase patient satisfaction (Kumar et al. 1997).

Optimal Scheduling and Sequencing Rules. We show that the robust appointment scheduling and sequencing problem with patient-specific maximum waiting time guarantees is $\mathcal{NP}$-hard in general, and introduce a mixed-integer linear programming formulation. We present optimal scheduling and sequencing rules for the following polynomial-time solvable problem variants. First, we consider the case of constant idle time costs, minimizing total idle time plus overtime cost. For this case, we prove optimality of an extended variant of the well-known Smallest-Variance-First rule, sorting patients in ascending order by the degree of uncertainty in their service time and their waiting time guarantees. For the case of non-increasing idle time costs and any given but fixed patient sequence, we prove optimality of the so-called ASAP rule, scheduling patients as early as possible while fulfilling the waiting time guarantees.

Computational Study Using Patient Data from a Radiology Department. Using real data from the radiology department of a large hospital, we first show that the problem can be solved in reasonable computation times for all considered instances. Second, we conduct a sensitivity analysis of the waiting times, idle times, and overtime obtained with optimal schedules and sequences. Third, we compare the schedules with those obtained by a weighted sum approach. The results show that waiting time guarantees lead to limited and more balanced waiting times across all patients. Moreover, by definition, our approach introduces the waiting time guarantees with a minimal increase in worst-case total cost. In particular, in cases in which the waiting times obtained by the weighted sum approach are also acceptable for all patients, our approach generally leads to lower total costs without increasing the waiting time of the longest waiting patient.

The remainder of the chapter is organized as follows. Section 2.2 reviews related optimization problems in the appointment scheduling literature. Section 2.3 formally introduces the Robust Appointment Scheduling Problem with Waiting Time Guarantees. We show that it is $\mathcal{NP}$-hard in general and derive a mixed-

integer linear program formulation. In Section 2.4, we present optimal scheduling and sequencing rules for relevant special cases in polynomial time. In Section 2.5, we conduct a computational study with real-world data, showing that introducing maximum waiting time guarantees can improve both waiting times and total cost compared to a weighted sum approach.

2.2. Literature Review

For an overview of general appointment scheduling, we refer the interested reader to Cayirli and Veral (2003), Gupta and Denton (2008), and Ahmadi-Javid et al. (2017). In the following, we review two streams of the appointment scheduling literature that are most relevant to our research. In Section 2.2.1, we discuss robust optimization approaches using box uncertainty sets for service times. In Section 2.2.2, we survey alternative approaches for modeling waiting time, distinct from the widely used weighted sum approach that minimizes total idle time, total waiting time, and overtime.

2.2.1. Robust Appointment Scheduling

Robust approaches to appointment scheduling that require only the shortest and longest service times but no other distributional information are scarce. To the best of our knowledge, Mittal et al. (2014) present the first such approach, minimizing a weighted sum of idle time costs and waiting time costs. Schulz and Udwani (2019) generalize the problem in Mittal et al. (2014) to idle time costs that are heterogeneous in the patient sequence and thus implicitly heterogeneous over time. Both works derive scheduling and sequencing rules for various special cases that are remarkably easy to implement. Given a patient sequence and homogeneous idle time costs, Mittal et al. (2014) present a polynomial-time algorithm to compute optimal interappointment times, which decrease as the planning horizon progresses. They also show that a simple variant of Smith's rule, scheduling patients in ascending order of the length of their service time intervals divided by their waiting time costs, is a 1.137 approximation. Schulz and Udwani (2019) prove a slightly adapted variant of this rule to be 1.0604-approximate, and the problem of simultaneously scheduling and sequencing patients with hetero-

geneous idle time costs and the weighted sum objective function to be strongly $\mathcal{NP}$-hard. Recently, Issabakhsh et al. (2021) addressed a robust appointment scheduling problem for infusion centers and solved it with a scheduling heuristic based on Adaptive Large Neighborhood Search. Advancing this first stream of literature, whose robust optimization framework we follow, we provide waiting time guarantees to all patients.

2.2.2. Modeling Waiting Time

Building on the classification by Bertsimas et al. (2011), we divide approaches for modeling waiting times into two substreams, based on the criterion used to evaluate waiting time. The utilitarian criterion minimizes a weighted sum of waiting times, while fairness criteria address individual waiting times more directly.

Utilitarian Criterion. In the first substream, we review appointment scheduling approaches that deviate from the classic weighted sum objective function but retain the utilitarian criterion. Sang et al. (2021) use a stochastic optimization approach to minimize any quantile of the total (idle time, overtime, and waiting) cost for a given job sequence. They find a semi-dome-shaped pattern with increasing interappointment times at the beginning of the planning horizon and constant interappointment times later. Zhou et al. (2022) use a chance-constraint approach, minimizing, among others, the probability that the sum of waiting times exceeds a certain threshold. They find optimal interappointment times to follow the well-known dome-shaped pattern in the case of zero no-shows and walk-ins. With a positive number of walk-ins, they find optimal interappointment times in the middle of the planning horizon to be constant, forming a so-called plateau dome-shaped pattern.

Fairness Criteria. Qi (2017) presents the first optimization study with the primary objective of fair patient waiting times. She lexicographically minimizes the so-called delay unpleasantness of the longest-waiting patient, which considers the frequency and intensity of delays above a certain threshold. Optimal scheduling rules depend on waiting time thresholds and the weighting of overtime costs. As

the weighting of overtime increases, the interappointment times between the first appointments are shorter than with weighted sum approaches. As the waiting time threshold decreases, later start times are assigned to the later appointments in the planning horizon. Following a queuing approach, Benjaafar et al. (2023), by contrast, assign the earliest possible appointment time to each patient request while limiting the expected waiting time of each patient. They find that the optimal times between appointments increase in the order of arrival. Finally, for a given patient sequence, Bendavid et al. (2018) minimize the completion time of the last appointment while bounding the probability of delay for each appointment. Under the assumption that the distribution of service times is known, they develop a simulation-based sequential algorithm and show that optimal interappointment times follow a dome-shaped pattern.

Our Contribution. In this work, we contribute to the fairness criteria by providing an a priori waiting time guarantee for each patient, effectively limiting their worst-case waiting time. These waiting time guarantees can be customized to match the tolerance levels for waiting identified in empirical studies. Moreover, compared to existing approaches, we allow for substantial cost reduction by accounting for (heterogeneous) idle time costs and the sequencing decision.

2.3. Problem Formulation

This section is organized as follows. In Section 2.3.1, we formally introduce the Robust Appointment Scheduling Problem with Waiting Time Guarantees. In Section 2.3.2, we show that the problem is $\mathcal{NP}$-hard. In Section 2.3.3, we derive a mixed-integer linear program.

2.3.1. Mathematical Model

We consider a set of n preselected patients denoted by $[n] := \{1, \ldots, n\}$, and a single service provider with a regular working time of L time units. Our decisions involve determining the patient sequence and appointment times. To determine the patient sequence, we introduce binary variables π_{ij} for $i, j \in [n]$, where $\pi_{ij} = 1$ if and only if exactly $i - 1$ patients are served before patient j. In other words,

patient j is assigned to the i-th appointment. We define continuous variables $A_i \geq 0$ to represent the start time of each appointment $i \in [n]$. We consider the following costs. Idle time before the first appointment incurs a per-time unit cost c_1, and idle time between two consecutive appointments $i-1$ and i incurs a per-time unit cost c_i, for $2 \leq i \leq n$. If there is any idle time before the planning horizon of L time units ends, it incurs a per-time unit cost c_{n+1}. Additionally, overtime incurs a per-time unit cost c_o. We aim to determine a patient sequence and appointment times that minimize the cost incurred by idle time and overtime, while ensuring that no patient $j \in [n]$ waits longer than their maximum waiting time guarantee W_j.

The main difficulty is that patients have uncertain service times. We assume that patients who show up are on time. For any given but fixed service time scenario $p \in \mathcal{P}$, sequence π, and appointment times A, we compute waiting times, idle times, and overtime as follows. We start by initializing $C_0(\pi, A, p) := 0$, and define completion times recursively for each appointment $i \in [n]$ as $C_i(\pi, A, p) := \max\{A_i, C_{i-1}(\pi, A, p)\} + \sum_{j=1}^{n} \pi_{ij} p_j$. It is useful and easy to see by induction that the completion times can be reformulated as $C_i(\pi, A, p) = \max_{l \in [1,i]} \left\{ A_l + \sum_{k=l}^{i} \sum_{j=1}^{n} \pi_{kj} p_j \right\}$, see Lemma 3 in Appendix A.1.1. The waiting time before each appointment $i \in [n]$ then equals the duration between the completion of the previous appointment and the start of the current appointment, expressed as

$$\text{wait}_i(\pi, A, p) := \left(C_{i-1}(\pi, A, p) - A_i\right)^+, \tag{2.1}$$

with $(\cdot)^+ := \max\{0, \cdot\}$. Similarly, the idle time before each appointment $i \in [n+1]$, with $A_{n+1} := L$, is given by $\left(A_i - C_{i-1}(\pi, A, p)\right)^+$, and overtime is computed as $\left(C_n(\pi, A, p) - L\right)^+$. The cost function

$$\begin{aligned} \text{cost}(\pi, A, p) := &\sum_{i=1}^{n} c_i \left(A_i - C_{i-1}(\pi, A, p)\right)^+ \\ &+ \max\left\{c_{n+1}(L - C_n(\pi, A, p)), c_o(C_n(\pi, A, p) - L)\right\} \end{aligned} \tag{2.2}$$

accounts for the total cost arising from idle time and overtime. We refer to Appendix B.2 for a summary of the notation.

To establish sustainable solutions, appointment schedules should be designed sufficiently robust to withstand unfortunate service times and no-shows. *Excessive* waiting times, idle time, and overtime not only incur high costs but also lead to employee and patient stress, dissatisfaction, and reduced loyalty. In healthcare, these factors can even harm patient health (Bae and Fabry 2014). Hence, we introduce a robust optimization approach that ensures smooth operations for all stakeholders, even in adverse scenarios. This gives rise to the following rbust appointment scheduling problem with waiting time guarantees:

$$\min \left\{ \max_{p \in \mathcal{P}} \{\text{cost}(\pi, A, p)\} \right\} \tag{2.3}$$

$$\text{s.t.} \max_{p \in \mathcal{P}} \left\{ \max_{i \in [n]} \left\{ \text{wait}_i(\pi, A, p) - \sum_{j=1}^{n} \pi_{ij} W_j \right\} \right\} \leq 0 \tag{2.4}$$

$$\sum_{i=1}^{n} \pi_{ij} = 1, \sum_{j=1}^{n} \pi_{ij} = 1 \qquad \forall i, j \in [n] \tag{2.5}$$

$$A_1 = 0 \tag{2.6}$$

$$\pi_{ij} \in \{0, 1\}, A_i \geq 0 \qquad \forall i, j \in [n] \tag{2.7}$$

Objective Function (2.3) minimizes the worst-case total cost incurred by idle time and overtime, preventing excessive idle time and excessive overtime. Constraint (2.4) ensures that the waiting time guarantees of all patients are met across all scenarios, ensuring acceptable waiting times. Thus, by preparing both the provider and the patients for their respective worst-case scenarios, Objective Function (2.3) and Constraint (2.4) ensure smooth operation even on a challenging day. Constraints (2.5) enforce that each patient is assigned to exactly one appointment, and each appointment is assigned to exactly one patient. Constraint (2.6) sets the start time of the first appointment to 0.

Remark 1. Due to the adversarial problems in Objective Function (2.3) and Constraint (2.4), which maximize the total cost and the waiting time of the longest-waiting patient, the above problem formulation is non-linear. In particular, it is not amenable to standard MILP solvers in this form.

Remark 2. The problem always has a feasible solution. For an arbitrary but

fixed sequence π, it is easy to see that the appointment times $A_1 := 0$ and $A_i := \left(\max_{l\in[1,i-1]} \left(A_l + \sum_{k=l}^{i-1}\sum_{j=1}^{n} \pi_{kj}\bar{p}_j\right) - \sum_{j=1}^{n} \pi_{ij}W_j\right)^+$ for $i \in [2,n]$ are feasible solutions. In the special case that all waiting time guarantees are so long that the latter term equals 0 for all $i \in [n]$, all patients could be scheduled at time 0, which corresponds to the single-block rule in Cayirli and Veral (2003).

Remark 3. It is noteworthy that monotonic increase is not guaranteed in the special case that $\max\{0, C_{i-1}(\pi, A, p) - A_{i-1}\} + \sum_{j=1}^{n} \pi_{i-1,j}p_j < \sum_{j=1}^{n} \pi_{ij}W_j$ holds for any appointment $i \in [n]$.

2.3.2. Complexity Analysis

In this section, we establish the $\mathcal{NP}$-hardness of the Robust Appointment Scheduling Problem with Waiting Time Guarantees, which holds even if all patients have identical waiting time guarantees. This result expands upon existing $\mathcal{NP}$-hardness proofs for appointment scheduling and sequencing problems employing weighted sum approaches (Bosch 1997, Kong et al. 2016, Schulz and Udwani 2019).

Theorem 1 ($\mathcal{NP}$-Hardness)**.** *Let all patients have identical waiting time guarantees W. Moreover, let the idle time costs fulfill $c_1 = \ldots = c_\sigma > c_{\sigma+1} = \ldots = c_n$ for some $\sigma \in [n-1]$ and the overtime cost $c_o = 0$. Then the Robust Appointment Scheduling Problem with Waiting Time Guarantees is $\mathcal{NP}$-hard.*

We briefly sketch the proof and refer to the formal proof in Appendix A.1.1 for details. In a worst-case scenario, worst-case idle times arise in two ways: If no idle time occurs up to and including appointment $i^* - 1$, then the idle time before appointment i^* equals $\sum_{i=1}^{i^*-1}\sum_{j=1}^{n} \pi_{ij}(\bar{p}_j - \underline{p}_j) - W$. The idle time before all subsequent appointments $i > i^*$ equals $\sum_{j=1}^{n} \pi_{i-1,j}(\bar{p}_j - \underline{p}_j)$. In the case of $i^* = \sigma$, the unit idle time cost before appointment i^* exceeds all later unit costs of idle time. The challenge lies in partitioning the customers into two groups such that the idle time sum $\sum_{i=1}^{i^*-1}\sum_{j=1}^{n} \pi_{ij}(\bar{p}_j - \underline{p}_j) - W$ is minimized. In Appendix A.1.1, we show that the well-known Subset Sum Problem can be reduced to this particular problem, establishing the $\mathcal{NP}$-hardness of the Robust Appointment Scheduling Problem with Waiting Time Guarantees.

Remark 4. If the assumption on idle time costs in Theorem 1 is replaced by constant idle time costs $c_1 = \ldots = c_n$, indicating no time preference for idle time occurrence, the Robust Appointment Scheduling Problem with Waiting Time Guarantees becomes solvable in polynomial time. We discuss this case in detail in Section 2.4.

2.3.3. Linearization

Due to the recursive nature of completion times in the cost function, applying the classical dualization scheme by Bertsimas and Sim (2004) for linearization is impractical. However, we overcome this challenge by leveraging structural properties to determine worst-case service times analytically. Specifically, we derive worst-case service times maximizing cost and waiting times, respectively. As a result, we present a compact mixed-integer linear program (MILP). For proofs not included in the main part, we refer the interested reader to Appendix 3.3.2.2.

Adversarial Problem of Maximizing Cost. The following theorem shows that the total cost is maximized in one of the $n+1$ sequence-dependent scenarios, in which patients scheduled for the first $i^* \in \{0, \ldots, n\}$ appointments require minimum service duration (idle time maximization), and all subsequent patients require maximum service time (overtime maximization).

Theorem 2. *Given a patient sequence π and appointment times A, let $\underline{C}_i(\pi, A) := \max\{A_i, \underline{C}_{i-1}(\pi, A)\} + \sum_{j=1}^{n} \pi_{ij}\underline{p}_j$ for $i \in [n]$, with $\underline{C}_0(\pi, A) := 0$, define the completion times when all patients require minimum service time.*

Then $\max_{p \in \mathcal{P}} cost(p, \pi, A)$ is equivalent to

$$\max_{i^* \in \{0,\ldots,n+1\}} \sum_{i=1}^{i^*} c_i (A_i - \underline{C}_{i-1}(\pi, A))^+ + c_o \left(A_{i^*} + \sum_{k=i^*}^{n} \sum_{j=1}^{n} \pi_{kj} \bar{p}_j - L \right)^+ \tag{2.8}$$

with $A_{n+1} := L$.

Proof. We show that any service time scenario that maximizes total cost can be decomposed into two parts: the first one maximizing idle time and the second one maximizing overtime. Assume w.l.o.g. that $\pi = id$.

First, using the reformulation of the completion times from Lemma 3 and $A_1 = 0$, we find:

$$\sum_{i=1}^{n+1} c_i \left(A_i - C_{i-1}(\pi, A, p)\right)^+ + c_o \left(C_n(\pi, A, p) - A_{n+1}\right)^+ \tag{2.9}$$

$$= \sum_{i=1}^{n} c_i \left(A_i - C_{i-1}(\pi, A, p)\right)^+ + \max\left\{c_{n+1}\left(A_{n+1} - C_n(\pi, A, p)\right), c_o(C_n(\pi, A, p) - A_{n+1})\right\} \tag{2.10}$$

$$= \sum_{i=2}^{n} c_i \left(A_i - \max_{l\in[i-1]} \left\{A_l + \sum_{k=l}^{i-1} p_k\right\}\right)^+ + c_{n+1}\left(A_{n+1} - \max_{l\in[n]} \left\{A_l + \sum_{k=l}^{n} p_k\right\}\right)^+ + c_o \left(\max_{l\in[n]} \left\{A_l + \sum_{k=l}^{n} p_k\right\} - A_{n+1}\right)^+ \tag{2.11}$$

Now, using $A_1 = 0$, let the service time scenario $p^{opt} \in \mathcal{P}$ maximize total cost (2.9). We distinguish the following two cases.

Case 1: $A_{n+1} - C_n(\pi, A, p^{opt}) \geq 0$.
In this case, overtime cost equals zero. Therefore, setting p_i^{opt} to $\underline{p}_i$ for each $i \in [n]$ will lead to an equal or higher value of (2.11).

Case 2: $C_n(\pi, A, p^{opt}) - A_{n+1} \geq 0$.
In this case, there are non-negative overtime costs. Let $l^* \in [n]$ chosen minimally such that $C_n(\pi, A, p^{opt}) = A_{l^*} + \sum_{k=l^*}^{n} p_k^{opt}$. In particular, it is easy to see that idle time cannot occur after the l^*th appointment. Then setting p_i^{opt} to $\bar{p}_i$ for all $i \geq l^*$ and to $\underline{p}_i$ for all $i \leq l^* - 1$ will again lead to an equal or higher value of (2.11). □

Adversarial Problem of Maximizing Waiting Times. As for each service time scenario $p \in \mathcal{P}$, and patient $j \in [n]$, the realized service time p_j is non-negative,

it is straightforward to see that the scenario of maximum service times

$$\bar{p} := (\bar{p}_1, \ldots, \bar{p}_n) \in_{p \in \mathcal{P}} \left\{ \max_{l \in [1,i-1]} \left(A_l + \sum_{k=l}^{i-1} \sum_{j=1}^{n} \pi_{kj} p_j \right) \right\} \tag{2.12}$$

uniformly maximizes the waiting time for each appointment $i \in [n]$ in Constraint (2.4).

MILP. Now, let Δ_i denote the idle time before appointment $i \in [n+1]$ and σ_{i^*} denote the overtime accumulated starting from appointment $i^* \in [n+1]$. Then the worst-case service time scenarios from Theorem 2 allow us to obtain the following mixed-integer linear program for the Robust Appointment Scheduling Problem with Waiting Time Guarantees (RASWTG):

$$\min U \tag{2.13}$$

$$\text{s.t. } U \geq \sum_{i=1}^{i^*} c_i \Delta_i + c_o \sigma_{i^*} \qquad \forall i^* \in [n+1] \tag{2.14}$$

$$\Delta_i \geq A_i - \underline{C}_{i-1} \qquad \forall i \in [n+1] \tag{2.15}$$

$$\underline{C}_i = \max\{A_i, \underline{C}_{i-1}\} + \sum_{j=1}^{n} \pi_{ij} \underline{p}_j \qquad \forall i \in [n] \tag{2.16}$$

$$\sigma_{i^*} \geq A_{i^*} + \sum_{k=i^*}^{n} \sum_{j=1}^{n} \pi_{kj} \bar{p}_j - L \qquad \forall i^* \in [n] \tag{2.17}$$

$$A_l + \sum_{k=l}^{i-1} \sum_{j=1}^{n} \pi_{kj} \bar{p}_j - A_i \leq \sum_{j=1}^{n} \pi_{ij} W_j \qquad \forall i, l : i \in [n], l \in [i-1] \tag{2.18}$$

$$\sum_{j=1}^{n} \pi_{ij} = 1, \sum_{i=1}^{n} \pi_{ij} = 1 \qquad \forall i, j \in [n] \tag{2.19}$$

$$A_1 = 0, \underline{C}_0 = 0, A_{n+1} = L \tag{2.20}$$

$$A_i, \underline{C}_i \geq 0 \qquad \forall i \in [n] \tag{2.21}$$

$$\Delta_i, \sigma_i \geq 0 \qquad \forall i \in [n+1] \tag{2.22}$$

$$\pi_{ij} \in \{0, 1\} \qquad \forall i, j \in [n] \tag{2.23}$$

Compared to (2.3) – (2.7), this MILP can be characterized as follows: Objective Function (2.13) minimizes the upper bound U set on the worst-case total cost incurred by idle time and overtime. Constraints (2.15) – (2.17), based on Theorem 2, determine the worst-case total cost. Constraint (2.18) ensures that the waiting time guarantees are fulfilled even in the worst case, using the waiting time reformulation (2.12).

RASWTG is an assignment problem of patients to appointments with assignment-dependent cost. It has $O(n^2)$ binary variables, $O(n)$ continuous variables, and $O(n^2)$ constraints. Using standard linearization techniques, Constraint (2.16) can be linearized by introducing another $O(n)$ binary variables. The computational study in Section 2.5 shows that RASWTG is solvable with a state-of-the-art MILP solver in short times for all instances with up to twenty patients from a data set of CT examinations in a large university hospital.

2.4. Optimal Scheduling and Sequencing

In this section, we present optimal scheduling and sequencing rules to solve RASWTG in polynomial time in special cases, as summarized in Table 2.1. Additionally, we investigate optimal interappointment times for the general case. All results demonstrate how to provide maximum waiting time guarantees cost-efficiently.

Table 2.1. Problem variants and exact solution methods

	Problem variant	Solution method
Section 2.4.1:	Constant idle time cost	SVF-WTG sequencing rule
Section 2.4.2:	Non-increasing idle time cost	ASAP scheduling rule
Section 2.4.3:	General case	MILP

2.4.1. Polynomial-Time Sequencing Rule for the Case of Constant Idle Time Costs and Zero No-Shows

Finding an optimal patient sequence can substantially reduce idle time costs (Cayirli et al. 2008, Deceuninck et al. 2018) but appears extremely difficult (Mak et al. 2015). Similarly, Theorem 1 implies that (unless $\mathcal{P} = \mathcal{NP}$) we cannot find a universal optimal scheduling and sequencing rule for the Robust Appointment Scheduling Problem with Waiting Time Guarantees in polynomial time. However, for a relevant special case, we now prove optimality of a polynomial-time sequencing rule consistent with the well-known Smallest-Variance-First (SVF) heuristic.

SVF is the most common heuristic for weighted sum approaches to appointment sequencing with stochastic service times. It ranks patients in ascending order of variance, following the intuition to reduce the probability of deviating from the schedule by placing patients with less variable service times first. For recent SVF studies in a stochastic optimization framework, we refer to Kong et al. (2016), Jafarnia-Jahromi and Jain (2020), and de Kemp et al. (2021). Moreover, Mak et al. (2015) prove, under a mild condition, the optimality of the SVF rule for a robust mean-variance model with an arbitrary number of appointments. Recently, Kong et al. (2016) showed that the optimality of the SVF policy depends on the number of patients as well as the shape of service time distributions.

Theorem 3 (Optimal Patient Sequence)**.** *Let idle time costs $c_1 = \ldots = c_{n+1}$ be constant, indicating no time preference for the occurrence of idle time. Then it is optimal to schedule patients $j \in [n]$ in non-decreasing order w.r.t. $\bar{p}_j - \underline{p}_j + (1 + c_o)W_j$.*

Under the given assumptions, it is optimal to schedule patients $j \in [n]$ in non-decreasing order with respect to their service time uncertainty $\bar{p}_j - \underline{p}_j$ and their waiting time guarantee W_j, as a function of overtime cost c_o. This result intuitively extends the SVF rule to account for waiting time guarantees: Uncertainty is kept as low as possible for as long as possible, particularly for patients with stricter waiting time guarantees. The relative importance of both criteria depends on the overtime cost. We call it the Smallest-Variance-First Rule with Waiting Time Guarantees (SVF-WTG).

Proof. Consider an optimal sequence π^* that does not adhere to the SVF-WTG sequencing rule. Then there exists $i', j', j'' \in [n]$ satisfying $\pi^*_{i',j'}, \pi^*_{i'+1,j''} = 1$ and $\bar{p}_{j'} - \underline{p}_{j'} + W_{j'}(1+c_o) > \bar{p}_{j''} - \underline{p}_{j''} + W_{j''}(1+c_o)$. To establish the optimality of the SVF-WTG sequencing rule, we prove that (iteratively) swapping the positions of patients j' and j'' does not increase the optimal worst-case cost $\max_{i^* \in \{0,\dots,n+1\}} \sum_{i=1}^{i^*} c_i \left(A_i - \underline{C}_{i-1}(\pi^*, A)\right)^+ + c_o \left(A_{i^*} + \sum_{i=i^*}^{n} \pi^*_{ij} \bar{p}_j - L\right)^+ =: opt.$

Since the idle time costs are constant and all patients show up, we know from a special case of Theorem 4 that $A_i = \left(\sum_{k=1}^{i-1} \sum_{j=1}^{n} \pi^*_{kj} \bar{p}_j - \sum_{j=1}^{n} \pi^*_{ij} W_j\right)^+$, with $A_1 = 0$, are optimal appointment times for $i \in [2, n]$.

First, we consider the costs associated with the case $i^* \in \{0, n+1\}$. For $i^* = n+1$, using the idle time cost reformulation (3.34) from Lemma 2, the total cost is equivalent to $c_{n+1}(L - \sum_{i=1}^{n} \sum_{j=1}^{n} \pi^*_{ij} \underline{p}_j)$, and swapping the position of two patients does not affect this value. For $i^* = 0$, where there is no idle time, swapping two patients does not affect the value.

For each $i^* \in [n]$, the associated cost is

$$\sum_{i=1}^{i^*} (A_i - \underline{C}_{i-1}(\pi^*, A))^+ + c_o \left(A_{i^*} + \sum_{k=i^*}^{n} \pi^*_{kj} \bar{p}_j - L \right) \tag{2.24}$$

Moreover, we assume w.l.o.g. that $A_{i^*} = \sum_{k=1}^{i^*-1} \sum_{j=1}^{n} \pi^*_{kj} \bar{p}_j - \sum_{j=1}^{n} \pi^*_{ij} W_j > 0$. Therefore, the total cost becomes

$$\sum_{i=1}^{i^*} (A_i - \underline{C}_{i-1}(\pi^*, A))^+ + c_o \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \pi^*_{ij} \bar{p}_j - \sum_{j=1}^{n} \pi_{i^* j} W_j - L \right) \tag{2.25}$$

Using again the idle time cost reformulation (3.34) from Lemma 2, the sequence-

dependent value is:

$$\sum_{i=1}^{i^*}(A_i - \underline{C}_{i-1}(\pi^*, A))^+ - c_o \sum_{j=1}^{n} \pi^*_{i^*j} W_j \tag{2.26}$$

$$= A_{i^*} - \sum_{k=1}^{i^*-1}\sum_{j=1}^{n} \pi^*_{kj}\underline{p}_j - c_o \sum_{j=1}^{n} \pi^*_{i^*j} W_j \tag{2.27}$$

$$= \sum_{k=1}^{i^*-1}\sum_{j=1}^{n} \pi^*_{kj}\bar{p}_j - \sum_{j=1}^{n} \pi^*_{i^*j} W_j - \sum_{k=1}^{i^*-1}\sum_{j=1}^{n} \pi^*_{kj}\underline{p}_j - c_o \sum_{j=1}^{n} \pi^*_{i^*j} W_j \tag{2.28}$$

$$= \sum_{k=1}^{i^*-1}\sum_{j=1}^{n} \pi^*_{kj}(\bar{p}_j - \underline{p}_j) - (1 + c_o) \sum_{j=1}^{n} \pi^*_{i^*j} W_j. \tag{2.29}$$

We now show that changing the sequence to $\pi_{i',j''}, \pi_{i'+1,j'} = 1$, the cost associated with each $i^* \in [n]$ does not exceed the optimal solution value obtained with π^*.

For each $i^* > i'+1$ or $i^* < i'$, the solution value does not change when changing the sequence to $\pi_{i',j''}, \pi_{i'+1,j'} = 1$. For $i^* = i'$, we find

$$\sum_{k=1}^{i^*-1}\sum_{j=1}^{n} \pi^*_{kj}(\bar{p}_j - \underline{p}_j) - (1 + c_o) \sum_{j=1}^{n} \pi^*_{i^*j} W_j \tag{2.30}$$

$$= \sum_{k=1}^{i^*-1}\sum_{j=1}^{n} \pi^*_{kj}(\bar{p}_j - \underline{p}_j) - (1 + c_o) W_{j'} \tag{2.31}$$

$$< \sum_{k=1}^{i^*-1}\sum_{j=1}^{n} \pi^*_{kj}(\bar{p}_j - \underline{p}_j) + \bar{p}_{j'} - \underline{p}_{j'} - (\bar{p}_{j''} - \underline{p}_{j''} + W_{j''}(1 + c_o)) \tag{2.32}$$

$$\leq \sum_{k=1}^{i^*}\sum_{j=1}^{n} \pi^*_{kj}(\bar{p}_j - \underline{p}_j) - W_{j''}(1 + c_o) \tag{2.33}$$

$$\leq opt - c_o \left(\sum_{j=1}^{n} \bar{p}_j - L \right) \tag{2.34}$$

and therefore, changing the sequence to $\pi_{i',j''}, \pi_{i'+1,j'} = 1$ does not increase the optimal solution value for the case of $i^* = i'$. Finally, for $i^* = i' + 1$, it is easy to

see that switching the position of j' and j'' improves the associated costs:

$$\sum_{k=1}^{i^*-2}\sum_{j=1}^{n}\pi^*_{kj}(\bar{p}_j-\underline{p}_j)+\bar{p}_{j'}-\underline{p}_{j'}-(1+c_o)W_{j''} \tag{2.35}$$

$$>\sum_{k=1}^{i^*-2}\sum_{j=1}^{n}\pi^*_{kj}(\bar{p}_j-\underline{p}_j)+\bar{p}_{j''}-\underline{p}_{j''}-(1+c_o)W_{j'} \tag{2.36}$$

We note that $i^* = i' + 1$ cannot have been the maximizer of the total worst-case cost, as this would contradict the optimality of π^*.

Thus, (iteratively) applying the SVF-WTG rule does not increase the total worst-case cost. □

2.4.2. Polynomial-Time Scheduling Rule under Non-Increasing Idle Time Costs

In the following, we assume that an arbitrary sequence of patients is given. We first present the so-called As Soon As Possible (ASAP) scheduling rule defined by Algorithm 1. Secondly, we show that in the case of non-increasing idle time costs, it leads to an optimal solution in polynomial time. Thirdly, we describe how ASAP generalizes the well-known Bailey-Welch rule (Welch and Bailey 1952).

Algorithm 1 ASAP scheduling rule

Input: $n \in \mathbb{N}, c \in \mathbb{R}^{n+1}$ with $c_1 \geq \ldots \geq c_{n+1}$, $(W_j)_{j\in[n]}$, $\bar{p} \in \mathbb{R}^n$, $k \geq 0$, $\pi \in \{0,1\}^{n\times n}$

1: $A_1 \leftarrow 0$ **for** $i \leftarrow 2$ *to* n **do**

2:
$$A_i \leftarrow \left(\max_{l\in[i-1]} A_l + \sum_{\sigma=l}^{i-1}\sum_{j=1}^{n}\pi_{\sigma j}\bar{p}_j - \sum_{j=1}^{n}\pi_{ij}W_j\right)^+$$

3:

4: **return** $(A_1, \ldots, A_n)$

The ASAP scheduling rule schedules each patient as early as possible while respecting the waiting time constraints. It is constructed as follows by Algorithm 1. The start time of the first appointment is set to 0 in Line 1. For each subsequent appointment, the algorithm computes the latest possible completion time of the previous appointment in Line 3 and subtracts the waiting time guarantee of the

patient scheduled for the i-th appointment. Notably, the appointment times become $A_i = \left(\sum_{k=1}^{i-1} \sum_{j=1}^{n} \pi_{kj} \bar{p}_j - \sum_{j=1}^{n} \pi_{ij} W_j \right)^+$.

Algorithm 1 computes optimal appointment times for a given sequence under non-increasing idle time costs, as stated in the following theorem and proven in Appendix 2.4.2. This condition, while crucial for the optimality proof, holds practical significance. Constant idle time costs indicate no time preference for the occurrence of idle time. Decreasing idle time costs indicate that later idle time is less expensive than earlier idle time, potentially due to its productive use for documentation work incurred by earlier appointments.

Theorem 4 (Optimal Appointment Times)**.** *Given non-increasing idle time costs $c_1 \geq \ldots \geq c_{n+1}$, and a patient sequence π, Algorithm 1 computes optimal appointment times in polynomial time for the Robust Appointment Scheduling Problem with Waiting Time Guarantees.*

Proof. It is straightforward to see that Algorithm 1 runs in polynomial time. Moreover, the waiting time constraints enforce

$$A_i \geq \left(\max_{l \in [i-1]} A_l + \sum_{\sigma=l}^{i-1} \sum_{j=1}^{n} \pi_{\sigma j} \bar{p}_j - \sum_{j=1}^{n} \pi_{ij} W_j \right)^+ \tag{2.37}$$

for all $i \in [2, n]$. We now show that appointment times, with $A_1 = 0$, which fulfill the inequality (2.37) with equality are an optimal solution in the case of non-increasing idle time costs.

Let A^* be optimal appointment times, which fulfill, for some $i' \in [2, n]$, that $A^*_{i'} > \left(\max_{l \in [i-1]} A^*_l + \sum_{\sigma=l}^{i-1} \sum_{j=1}^{n} \pi_{\sigma j} \bar{p}_j - \sum_{j=1}^{n} \pi_{ij} W_j \right)^+$. As $c_{i-1} - c_i \geq 0$ for all $i \in [2, n]$, and using the cost reformulation (3.31) – (3.34) in the proof of Lemma 2 with corresponding index set $\mathcal{I}$, we distinguish the following two cases. Firstly, if $i' \in \mathcal{I}$, then choosing

$$A^*_{i'} = \left(\max_{l \in [i-1]} A^*_l + \sum_{\sigma=l}^{i-1} \sum_{j=1}^{n} \pi_{\sigma j} \bar{p}_j - \sum_{j=1}^{n} \pi_{ij} W_j \right)^+ \tag{2.38}$$

leads to a better objective value, contradicting the optimality of A^*. Secondly, if $i' \notin \mathcal{I}$, then choosing (2.38) does not change the optimal solution value. Thus, it

is optimal to schedule each appointment time A_i, $i \in [n]$, as early as possible. □

The ASAP rule can be considered a generalization of Bailey's rule, which schedules the first two patients simultaneously and all subsequent patients sequentially. In Line 3, each patient is scheduled to arrive at time 0 as long as the latest possible completion times of the previous patients are not longer than their waiting time guarantee. If this condition is no longer met, patients are scheduled sequentially. In particular, the maximum waiting time guarantees control how many patients can be scheduled at the start time of the planning horizon.

2.4.3. Optimal Interappointment Times for Varying Costs

In the literature, empirical studies on optimal interappointment times are commonly pursued in settings where analytical approaches face significant challenges. Accordingly, while we have proven optimality of the ASAP scheduling rule for the case of non-increasing costs, we derive further insights for the case of increasing costs by analyzing optimal interappointment times. The optimal schedules are computed using Gurobi and illustrated by the example of ten identical patients, each with a minimum service time of 15 minutes, a maximum service time of 25 minutes, and a waiting time guarantee of 30 minutes.

In the cases of non-increasing idle time costs, we observe that the optimal interappointment times follow a variant of the well-known plateau dome shape (Denton and Gupta 2003, Robinson and Chen 2003, Kong et al. 2020, Sang et al. 2021). Initially, interappointment times increase, followed by a plateau of constant interappointment times. The optimality of the ASAP scheduling rule implies increasing or varying interappointment times at the beginning of the scheduling

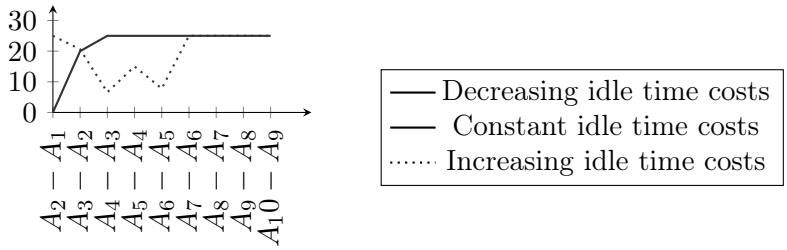

Figure 2.1. Optimal interappointment times (in minutes)

horizon to avoid early idle times, followed by a constant pattern limited by the waiting time guarantees. In this sense, the observed semi-plateau-dome pattern represents a customer-driven variant of the plateau-dome-shaped pattern.

On the other hand, the complexity observed in the case of increasing costs suggests that establishing optimal scheduling rules in this case may be more difficult. In this case, a balance must be found between the following strategies. On the one hand, scheduling the last appointment as early as possible is still essential to avoid overtime costs. On the other hand, the higher cost of later idle times can only be prevented by allowing for relatively long (worst-case) waiting times for later appointments. Achieving this while meeting waiting time guarantees may require scheduling early appointments later than strictly necessary. This may also lead to the appointment time of the last appointment being later than required by the waiting time constraints.

2.5. Computational Study

In this section, we evaluate the performance of RASWTG schedules on a real-world data set from the radiology department of a large university hospital. In particular, we analyze runtimes, conduct a sensitivity analysis, and present managerial insights from comparing waiting times, idle times, and overtime obtained with RASWTG to those obtained with a weighted sum approach.

2.5.1. Instances

We generate test instances based on 7,745 CT examination records performed on one of three CT machines in the radiology department of a large university hospital in Munich, Germany. All exams were performed in 2019 or the first quarter of 2020. For each CT examination, we have the following information: Exam type, exam day, start time, and completion time. The start and completion times were recorded as so-called "Modality Performed Procedure Steps". That is, they do not include setup times before and after the irradiation procedure, such as preparing the patient at the unit or removing the patient from the unit, but only the scanning time. Service times are calculated as the difference between completion time and start time. Figure A.1 in the Appendix A.1.2.1 presents a

boxplot illustrating the service times for the most common examination types.

We first split the data into an interval estimation set and a test set. The interval estimation set consists of all 6,359 CT examinations recorded in 2019; the test set consists of all 1,386 CT examinations recorded in the first quarter of 2020. The interval estimation set is used to estimate the service time intervals for the examinations in the test set. Specifically, for each of the 64 examination types, we estimate the minimum and maximum service time as the 5th and 90th percentiles of the empirical distribution of service times within the interval estimation set. We tested that this choice ensures meeting the waiting time guarantees for nearly all patients while also resulting in reasonable total costs. We then compute schedules for the examination requests recorded in the test set using these service time interval estimates. In case of multiple optimal solutions, we choose one that minimizes the sum of appointment times in accordance with the ASAP rule. The schedules are evaluated using the service times recorded in the test set.

To generate test instances, we consider three factors: number of patients, waiting time guarantees, and structure of the idle time costs. In a ceteris paribus design, we systematically vary these factors using three values each, as outlined in Table 2.2. Unless specified otherwise, our analysis centers on the base case characterized by bold values, encompassing constant idle time costs, ten patients, and a maximum waiting time guarantee of 30 minutes.

Table 2.2. Parameter sets (base case in bold)

Number of patients	Waiting time guarantee	Idle time cost
{5, **10**, 20}	{10 min, 20 min, **30 min**}	{**const.**, dec., inc.}

The chosen patient numbers, $n \in \{5, 10, 20\}$, are aligned with the daily CT examination records and consistent with previous computational studies utilizing distributionally robust optimization approaches (Qi 2017, Jiang et al. 2017, Kong et al. 2020). For each $n \in \{5, 10, 20\}$, patient selection is performed by first identifying days within the test set that contain at least n appointments. Subsequently, we randomly sample n consecutive patient records from these days. This process generates 252 instances across all classes with five patients, 107 instances across all classes with ten patients, and 39 instances across all classes with twenty

patients.

We differentiate three types of idle time costs: constant, decreasing, and increasing. Constant idle time costs $(c_i)_{i\in[n]}$ remain fixed at $c_i = 1$ for all $i \in [n+1]$. Decreasing idle time costs follow three assumptions: Idle time before the first appointment incurs a per-unit time cost of 1, idle time after the last appointment, fully utilized by staff for tasks like documentation, incurs a per-unit time cost of only 0.5, and costs linearly decrease between the first and last idle time. This results in $c_i = 1 - \frac{i-1}{2n}$ for idle time before appointment $i \in [n+1]$. For completeness, we also consider linearly increasing idle time costs with $c_i = \frac{n+i-1}{2n}$, $i \in [n+1]$. Overtime cost, complying with current US federal law requiring a minimum of 1.5 times the regular pay rate, is set at $c_o = 1.25$.

Finally, the length L of the planning horizon crucially impacts the trade-off between idle time cost and overtime cost. A short planning horizon may cause significant worst-case overtime, while an extended one may lead to no overtime but substantial idle time. For a comprehensive evaluation of the model, we select the length of the planning horizon for each instance to lie at a specific point between the earliest and latest possible completion times of the last appointment, accommodating both idle time and overtime. Specifically, L is chosen as the sum of maximum service times minus the waiting time guarantee for the last appointment. For instances with a waiting time guarantee of 30 minutes, the resulting average planning horizon is 50.08 minutes for five patients, 132.13 minutes for ten patients, and 312.53 minutes for twenty patients. The average cumulative uncertainty within this horizon, obtained as the sum of the service time interval lengths, amounts to 46.88 minutes for five patients, 91.03 minutes for ten patients, and 178.15 minutes for twenty patients.

2.5.2. Runtimes

In Table 2.3, we report the average runtimes to compute an optimal schedule using the polynomial-time algorithm (PTA) that results from first applying Theorem 3 and then Algorithm 1 in the case of constant idle time costs, and the mixed-integer program RASWTG otherwise. All experiments have been performed on an Intel Core i7-10510U 1.8 GHz machine (in 64-bit mode) with 16 GB memory under Microsoft Windows 10 and Gurobi 9.5.0 in Python 3.9.0.

Table 2.3. Average runtimes to compute an optimal schedule
(in seconds)

	Method		Patients		
Costs	RASWTG	PTA	5	10	20
Constant		•	0.00	0.00	0.00
Decreasing	•		0.01	0.45	197.35
Increasing	•		0.05	2.80	199.06

In the case of constant idle time costs, our polynomial-time algorithm efficiently computes optimal schedules in milliseconds, even for a substantial number of patients. When there are non-constant idle time costs, we employ the mixed-integer linear program RASWTG. With this model, Gurobi finds optimal schedules within seconds for up to 10 patients and within 200 seconds for 20 patients.

Furthermore, the observed runtimes support the conjecture from Section 2.4 that decreasing idle time costs lead to simpler optimal solution structures than increasing idle time costs.

2.5.3. Results

We first conduct a sensitivity analysis and then compare the results with those obtained by minimizing a weighted sum of waiting times, idle times, and overtime.

2.5.3.1. Sensitivity Analysis

We assess the impact of each of the three factors listed in Table 2.2 on waiting times, idle times, and overtime, from which we derive practical managerial recommendations. To comprehensively understand the dynamics, we recall the trade-offs between waiting time and idle time until the last appointment, and the trade-off between overtime and idle time following the last appointment.

Base Case. In the base case, the average waiting time per patient is 10.92 minutes, and the waiting time guarantees are met for 97% of the patients. The slight discrepancy from 100% is due to the estimation of minimum and maximum

Table 2.4. Number of patients (idle time and overtime in minutes)

	5 patients	10 patients	20 patients
Patients whose waiting time guarantee is met:	96.11%	96.88%	98.23%
Average idle time per appointment:	1.90	3.77	4.92
Average overtime:	7.92	1.77	0.22

service times derived from empirical service time data in the separate interval estimation set, excluding outliers. In cases where realized service times are outside the estimated range, the waiting time guarantees of a few later patients are unmet. Furthermore, the average total idle time amounts to 37.67 minutes and overtime to 1.77 minutes, relative to an average cumulative uncertainty of 91.03 minutes. Figure 2.2 illustrates the distribution of waiting and idle times by appointment.

Number of Patients. Table 2.4 summarizes our findings regarding waiting times, idle times, and overtime across a varying number of patients. Given our choice of the planning horizon L, overtime decreases as the number of patients rises. Moreover, as the number of patients in the planning horizon increases, the RASWTG approach generally leads to shorter waiting times and longer idle times. Specifically, the ASAP scheduling rule, which is optimal for constant and decreasing costs (see Theorem 4), provides a rationale for this effect:

The first few appointments can all be scheduled at the start time of the scheduling horizon, as long as the sum of their maximum service times does not exceed the waiting time guarantee of the subsequent appointment. For these appointments, waiting times accumulate in proportion to the service times realized. For all subsequent appointments, the start time can be calculated as the worst-case sum of the previous service times minus the waiting time guarantee. As the number of customers treated increases, we consider it increasingly unlikely that all customers require their maximum service time. Consequently, we expect short waiting times and longer idle times before late appointments.

We note that this effect could be mitigated by using uncertainty budgets that control the number of customers who require maximum service time.

Table 2.5. Average results by waiting time guarantee

	Idle	$\overline{\text{Wait}}$	WTG	Over
10-minute guarantee:	56.75	2.73	91.50%	0.85
20-minute guarantee:	47.02	6.01	94.86%	1.12
30-minute guarantee:	37.67	10.92	96.88%	1.77

idle: total idle time (in minutes), $\overline{\text{wait}}$: average waiting time (in minutes), WTG: percentage of patients for whom the waiting time guarantee is met, over: overtime (in minutes)

Waiting Time Guarantees. Practitioners should tailor waiting time guarantees to align with their patients' waiting time tolerance levels. Figure 2.2 illustrates how waiting and idle times vary across the planning horizon for different waiting time guarantee choices. Table 2.5 shows the corresponding total idle time, average waiting times, and adherence to the waiting time guarantee. Increasing the waiting time guarantee by 10 minutes, on average, improves the fulfillment of the waiting time guarantee by 1.90%, boosts capacity utilization by 18.44%, and maintains stability in terms of overtime. This substantial capacity utilization increase results from greater flexibility in terms of the waiting time constraints and, correspondingly, a shorter overall planning horizon.

Idle Time Costs. The idle time costs control whether idle times occur rather early or late in the planning horizon and, oppositely, whether waiting times occur rather early or late. Decreasing idle time costs, by definition, particularly avoid early idle times. Thus, idle times predominantly occur later in the planning

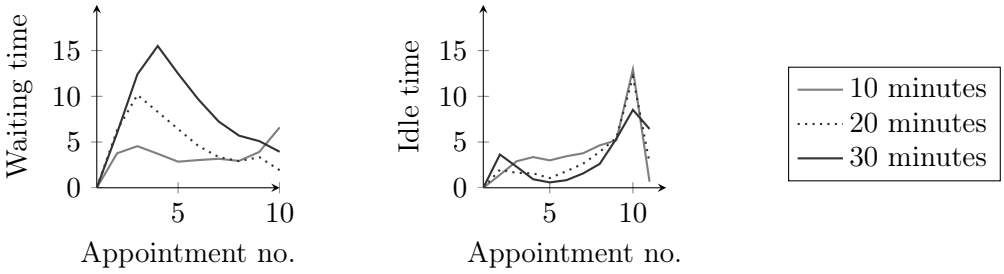

Figure 2.2. Average waiting and idle times by waiting time guarantee (in minutes)

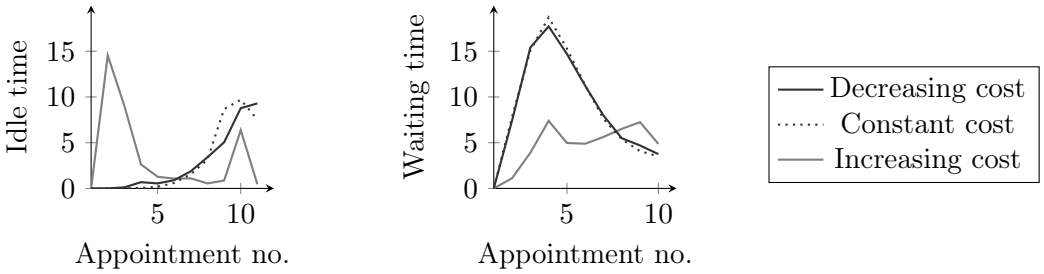

Figure 2.3. Average waiting and idle times by cost structure
(in minutes)

horizon, as can be seen in Figure 2.3 for the base case. The opposite is true for increasing idle time costs. Constant idle time costs do not define a preference for when idle time occurs. Notably, both total idle time and overtime are highest in the case of increasing idle time costs and significantly lower and closely aligned for constant and decreasing idle time costs. This observation suggests that the simple greedy rule which utilizes capacities as early as possible proves effective when minimizing total idle time. Furthermore, constant and decreasing idle time costs lead to later and, thus, potentially more productively usable idle times.

2.5.3.2. Comparison to a Weighted Sum Approach

In this section, we compare waiting times, idle times, and overtime acquired using RASWTG in contrast to those obtained using a weighted sum approach. Unless stated otherwise, we use the RASWTG approach in the base case setting. We refer to the following problem as Weighted Sum Robust Appointment Scheduling Problem (WSRAS).

$$\min_{\pi, A} U \tag{2.39}$$

$$\text{s.t. } U \geq \sum_{i=1}^{n+1} c_i \left(A_i - C_{i^*,i-1}\right)^+ + c_o \left(C_{i^*n} - L\right)^+ + \sum_{i=1}^{n} c_w \left(C_{i^*,i-1} - A_i\right)^+ \qquad \forall i^* \in [n+1] \tag{2.40}$$

$$C_{i^*i} = \max\{A_i, C_{i^*,i-1}\} + \sum_{j=1}^{n} \pi_{ij} \underline{p}_j, \qquad \forall i^* \in [n+1], i < i^* \tag{2.41}$$

$$C_{i^*i} = \max\{A_i, C_{i^*,i-1}\} + \sum_{j=1}^{n} \pi_{ij} \bar{p}_j, \qquad \forall i^* \in [n+1], i \geq i^* \tag{2.42}$$

$$\sum_{j=1}^{n} \pi_{ij} = 1, \sum_{i=1}^{n} \pi_{ij} = 1 \qquad \forall i, j \in [n] \tag{2.43}$$

$$A_1 = 0, A_{n+1} = L, C_{i^*0} = 0 \qquad \forall i^* \in [n+1] \tag{2.44}$$

Objective Function (2.39) minimizes the upper bound on total cost, as established in Constraint (2.40). Specifically, Constraint (2.40) upper bounds the sum of waiting times, weighted by c_w, idle times, weighted by sequence-dependent cost c_i, and overtime, weighted by c_o; while ensuring robustness with respect to the same worst-case scenarios as RASWTG (see Theorem 2). Constraints (2.41) and (2.42) define the completion times, while Constraint (2.43) represents the assignment constraint known from the Robust Appointment Scheduling Problem with Waiting Time Guarantees. To enhance our understanding of optimal WS-RAS schedules, we briefly explore optimal interappointment times, as illustrated in Figure A.2 in Appendix A.1.2.1.

Notably, Mittal et al. (2014) study a related robust appointment scheduling problem that minimizes a weighted sum of idle time and waiting time costs, under a collective worst-case scenario. Our works show, independently of each other and with different objective functions, that it is sufficient to consider the same $n + 1$ critical service time scenarios (see Theorem 2).

Waiting Times. We compare, first, how to control waiting times with both approaches and, second, the distribution of waiting times across all appointments. Applying the weighted sum approach, we need to carefully tune the waiting time costs to manage waiting times. In contrast, when applying RASWTG, the waiting time guarantees constitute input parameters which (a priori) limit all worst-case waiting times, and indirectly also the realized waiting times. It is therefore easier to control waiting times with RASWTG than with the weighted sum approach.

In addition, Figures 2.4 and 2.5 present boxplots illustrating patient waiting times depending on their position in the appointment schedule. These waiting times correspond to the optimal interappointment intervals depicted in Figure A.2 in the Online Appendix. Notably, for modest waiting time costs (c_w), waiting times with the weighted sum approach exhibit an increasing trend until the penultimate or last appointment, and excessive waiting times for some later appointments. High waiting time costs lead to very short waiting times for all appointments. In contrast, the RASWTG approach demonstrates an initial (limited) increase in waiting times followed by a gradual decrease. It is worth recalling that this decline can be mitigated by introducing uncertainty budgets. Furthermore, for comparable overall costs, waiting times are distributed more equitably across appointments with RASWTG, and excessive waiting times are less likely.

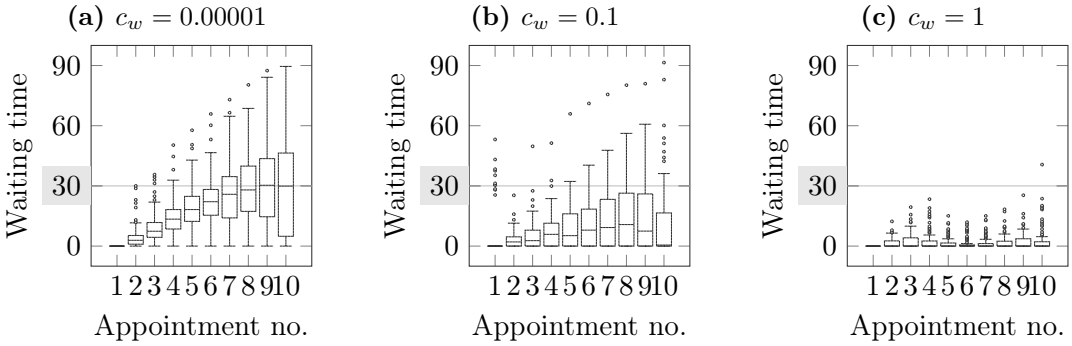

Figure 2.4. WSRAS waiting times by appointment position
(in minutes)

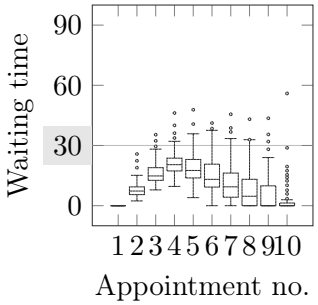

Figure 2.5. RASWTG waiting times by appointment position
(in minutes)

Costs. We finally evaluate the cost of replacing the weighted sum approach with RASWTG. We consider that a healthcare provider replaces WSRAS schedules with fixed waiting time cost (i) $c_w = 0.00001$, (ii) $c_w = 0.1$, and (iii) $c_w = 1$ over all days. Our findings are summarized in Table 2.6.

Let us first examine overtime. For WSRAS, we observe that overtime decreases with lower waiting time costs, as lower waiting costs allow for scheduling appointments earlier. Surprisingly, RASWTG exhibits the lowest overtime, on average only 1.77 minutes. This suggests that, even with hard waiting time constraints, RASWTG permits slightly earlier scheduling of the last appointment even compared to the weighted sum approach with low waiting time cost. Overall, both approaches maintain reasonable average overtime within a 1 to 7-minute range, and minimal with RASWTG.

Considering total idle time, we find that it naturally increases with rising waiting time costs for WSRAS. RASWTG displays the lowest total idle time, a result

Table 2.6. Comparison of WSRAS with RASWTG

	WSRAS $c_w = 0.00001$	WSRAS $c_w = 0.1$	**RASWTG**	WSRAS $c_w = 1$
Waiting time ≤ 30:	79.91%	92.43%	**96.88%**	99.91%
Idle time until A_n:	10.12	24.75	**17.18**	37.50
Total idle time:	37.92	39.03	**37.67**	42.27
Overtime:	2.17	3.13	**1.77**	6.37

consistent with our findings regarding overtime. It is worth noting that idle time for both approaches could be reduced by opting for a shorter planning horizon L but at the expense of increased overtime, which we aimed to prevent in the first place. However, as a substantial amount of idle time remains between the last appointment and the end of the planning horizon L for both approaches, we also compare the idle time up to the last appointment, which is less dependent on the choice of L. Again, for all waiting time costs except for $c_w = 0.00001$, which comes with excessive waiting times, RASWTG yields lower idle time than WSRAS.

We conclude with the following rule of thumb. In environments where most patients experience acceptable waiting times, introducing RASWTG offers both a priori waiting time guarantees and total cost reduction. The cost reduction typically comes with a limited increase in average waiting times. In environments where many patients experience excessive waiting times, introducing RASWTG effectively prevents excessive waiting times with a minimum increase in total worst-case cost. In all cases, RASWTG effectively mitigated excessive waiting times and overtime.

2.6. Conclusion

This work was motivated by the lack of intraday appointment scheduling and sequencing approaches to precisely control waiting times. We introduce maximum waiting time guarantees to respect patients' waiting time tolerance even in a worst-case scenario of uncertain service times. Hence, the established guarantees prevent negative effects of waiting times on patient dissatisfaction, perceived service quality, and loyalty shown in previous empirical research. At the same time, we minimize worst-case costs incurred by idle time and overtime.

We show that the presented appointment scheduling problem with waiting time guarantees is $\mathcal{NP}$-hard and find optimal scheduling and sequencing rules in polynomial time for practically relevant special cases. Assuming constant idle time costs, we show optimality of a sequencing rule that schedules patients in ascending order with the amount of uncertainty about their service times and their waiting tolerance. This rule provides a deeper understanding of the optimality of the well-known SVF rule and extends it to consider waiting time guarantees.

Moreover, for the case of non-increasing idle time costs, we show optimality of the ASAP scheduling rule that sets each appointment time as early as possible while fulfilling the waiting time guarantees. The presented sequencing and scheduling rules are intuitive and easy to implement.

Our case study with real data from the radiology department of a large hospital shows that Robust Appointment Scheduling Problem with Waiting Time Guarantees not only prevents excessive waiting times. Compared to weighted sum approaches, our approach also provides more flexibility to control the timing of idle and waiting times. Most importantly, we introduce the waiting time guarantees with a minimal increase in worst-case costs, which in many cases is even negative. Thus, compared to weighted sum approaches, waiting time guarantees may even reduce costs.

Our work also points to several avenues for future research. First, an intriguing open question involves introducing uncertainty budgets to control the level of conservatism, and adapting the presented scheduling and sequencing rules. Second, it would be interesting to integrate the consideration of no-shows into our robust framework. Third, the concept of minimizing costs under robust service levels is applicable to any service in which patient disutility increases sharply upon surpassing a critical threshold. Identifying such services and studying the corresponding problem variants seems promising.

3. Robust Predictive Overbooking

Co-authors: Rainer Kolisch and Andreas S. Schulz
Working paper.

Abstract. Overbooking is a widely adopted strategy to mitigate reduced capacity utilization stemming from missed appointments. Integrating individual no-show predictions into overbooking systems offers potential for optimizing capacity utilization without exacerbating waiting times or overtime. Nonetheless, ethical considerations have emerged concerning scheduling patients from higher-risk no-show groups for appointments associated with longer waiting times. In this study, we propose robust waiting time guarantees for all patients within the context of predictive overbooking. Building upon the cardinality-constrained robust optimization framework, we introduce uncertainty budgets for false positive and false negative predictions. Our approach ensures the fulfillment of waiting time guarantees under worst-case false predictions while minimizing total costs incurred by idle time and overtime. Leveraging a multi-layered decomposition approach and employing total unimodularity as well as LP duality, we formulate the problem as a mixed-integer linear program. Computational findings demonstrate consistent adherence to waiting time guarantees in over 96% of cases across all scenarios. Furthermore, total costs decrease naturally with prediction accuracy, with false positives exhibiting a slightly more adverse impact than false negatives. Interestingly, a higher no-show rate contributes to cost reduction.

3.1. Introduction

Missed appointments, known as no-shows, profoundly affect healthcare capacity utilization. Gier (2017) estimates the annual cost of no-shows at $150 billion in the United States alone. Data from the National Health Service NHS (2019) reveal approximately 307 million missed appointments in general practices annually, resulting in a cost exceeding £216 million and significantly hampering access to healthcare.

Overbooking strategies have been widely adopted to address the problem of reduced capacity due to no-shows, accepting more patients than available capacity. They have proven successful in improving capacity utilization and reducing costs, as evidenced by studies such as LaGanga and Lawrence (2007), Sickinger and Kolisch (2009), LaGanga and Lawrence (2012), Zheng et al. (2015), on the one hand. On the other hand, however, overbooking has been long known to incur excessive waiting times for some patients, if more patients than planned show up (Cayirli and Veral 2003, Gupta and Denton 2008, Dantas et al. 2018).

The recent surge in available data and predictive analytics methodologies may allow us to accurately forecast individual no-show candidates. Integrating individual no-show predictions into appointment scheduling frameworks - so-called predictive overbooking - holds the promise to optimize capacity utilization without increasing waiting times or overtime. A growing body of literature uses individual no-show predictions to find schedules that minimize a weighted sum of waiting times, idle times, and overtime. Yet, Samorani et al. (2021) demonstrated that if a group of patients has been characterized by a higher no-show risk in historical data, this objective function may result in patients in that group being disproportionately scheduled for appointments with longer wait times, emphasizing the need for more equitable modeling approaches.

To harness the potential of predictive overbooking for enhanced capacity utilization while mitigating its negative effects on waiting times, this work proposes a robust approach that accounts for uncertainty in the predictions. Instead of the classical approach of relying on no-show probabilities, we incorporate binary classification results for individual no-show predictions. Each patient's data includes their service time range and a binary individual no-show prediction. Building upon the cardinality-constrained robustness framework proposed by Bertsimas

and Sim (2004), we introduce uncertainty budgets Γ^{FP} and Γ^{FN} for false positive and false negative predictions, to the best of our knowledge a novel contribution to the field. Integrating these predictions and their robustness against imprecision, we extend the Robust Appointment Scheduling Problem with Waiting Time Guarantees, as introduced in Chapter 2, to include no-shows in this chapter. Specifically, we ensure meeting waiting time guarantees even if Γ^{FP} of predicted no-shows attend and Γ^{FN} of predicted show-ups do not, while minimizing total costs incurred by idle time and overtime.

The main contributions can be summarized as follows:

Robust Predictive Overbooking. Firstly, we directly take into account the results from a binary classification problem when predicting no-shows. Secondly, we correct the uncertainty with regard to the predictions by introducing uncertainty budgets for false positives and false negatives. Thirdly, we extend the problem setting of the Robust Appointment Scheduling Problem with Waiting Time Guarantees from Chapter 2 to account for no-shows. Thereby, we introduce maximum waiting time guarantees for (predictive) overbooking.

Mixed-Integer Linear Program. We employ a multi-layered decomposition approach to address the problem. Decomposing both the worst-case waiting time constraint, as well as the cost function, along with total unimodularity and LP duality, we linearize the problem. This results in a mixed-integer linear program for robust predictive overbooking with waiting time guarantees.

Computational Study. The results consistently show over 96% adherence to waiting time guarantees across all scenarios. While total costs naturally decrease with prediction accuracy, it is noteworthy that false positives appear to have a more adverse impact than false negatives. Moreover, intriguingly, we observe that a higher no-show rate even leads to cost reduction.

This chapter is structured as follows: Section 3.2 reviews related optimization problems in the literature on no-shows and appointment scheduling. Section 3.3 formally introduces the robust predictive overbooking problem with waiting time guarantees and presents a mixed-integer linear program formulation. In Sec-

tion 3.4, we conduct a computational study based on real-world data, showing that the consideration of no-shows can significantly increase capacity utilization without compromising the patients' waiting time guarantees.

3.2. Literature Review

This section explores predictive overbooking approaches in the context of appointment scheduling, building upon the extensive literature in the field (Cayirli and Veral 2003, Gupta and Denton 2008, Ahmadi-Javid et al. 2017), and the review on robust appointment scheduling in Section 2.2.1. To better understand individual no-show predictions, we first provide a brief overview of statistical methods and results on no-shows in Section 3.2.1. Subsequently, in Section 3.2.2, we review appointment scheduling systems that integrate these predictions.

3.2.1. No-Show Predictions

For a comprehensive literature review on no-show predictions in healthcare, we refer to the reviews by Deyo and Inui (1980), Bean and Talaga (1992), Garuda et al. (1998), Dantas et al. (2018), Carreras-García et al. (2020). Notably, Dantas et al. (2018) find an average no-show rate of 23% across all studies, with significant disparities observed by region and specialty, amongst others. They underscore the context-specific nature of no-show predictions. Here, we review the most important features, methods, and evaluation results.

Feature Selection. We follow Dantas et al. (2018) to categorize relevant features into appointment type, patient demographics, and patient behavior.
Appointment-related factors include lead time, specialty, appointment date/time, and referral source. Notably, lead time stands out as an important feature, with longer intervals between appointment scheduling and the appointment date correlating with higher rates of no-shows (Dantas et al. 2018, Salazar et al. 2022). Patient demographics such as age, gender, geographic proximity, and socioeconomic status, including insurance status, have yielded mixed findings (Dantas et al. 2018, Salazar et al. 2022, Liu et al. 2022). It is important to recognize that these characteristics entail sensitive information, thus requiring ethical and

cautious implementation. Finally, next to lead time, prior no-show history has emerged as the commonly reported most important feature to predict no-shows (Dantas et al. 2018, Chen et al. 2020, Liu et al. 2022).

Methods. Commonly approached as a binary classification task, predicting no-shows has been explored using a wide range of statistical learning methods. Historically, logistic regression has successfully prevailed (Carreras-García et al. 2020). Recent studies have shown decision tree algorithms outperforming logistic regression and competing with artificial neural networks (Fan et al. 2021, Salazar et al. 2022). Notably, in Chen et al. (2020), the XGBoost model outperformed existing models. Nevertheless, Nevertheless, it has been shown that logistic regression can even compete with artificial neural networks if previous no-show history is available as a significant feature (Liu et al. 2022). Clearly, the choice of method is context- and data-specific.

Evaluation Results. Alaeddini et al. (2011) achieved an 80% accuracy with a hybrid model of logistic regression and Naive Bayes. Mohammadi et al. (2018) reported accuracies of 73% for logistic regression, 71% for artificial neural networks, and 82% for Naive Bayes. In Batool et al. (2021), accuracy reached up to 95%, with a false positive rate of 7.39% with the lead time attribute and 41.81% without. Additionally, Liu et al. (2022) reported an accuracy of 83% in pediatrics, with a false positive rate of 17%. It is important to recognize that accuracy may overstate model performance due to data imbalance, and thus to consider other evaluation metrics such as false positive and false negative rates. While the results are promising, these findings also emphasize the need for individualized and context-specific studies of no-shows, particularly with the growing availability of data.

3.2.2. Predictive Overbooking

The literature on appointment scheduling under uncertain patient no-show behavior explores various approaches. For an extensive overview, we refer to Salah and Srinivas (2022). While some studies focus on integrating static no-show probabilities (Luo et al. 2012, Cayirli et al. 2012, Jiang et al. 2017), or time-dependent

no-show probabilities (LaGanga and Lawrence 2012, Kong et al. 2020), we provide a brief review of the growing body of appointment scheduling approaches that integrate individual no-show predictions.

Firstly, the structure of optimal appointment schedules with individual no-show probabilities has been examined when minimizing a weighted sum of waiting times, idle times, and overtime. Zacharias and Pinedo (2014) propose an overbooking approach that minimizes a weighted sum of waiting times, idle time, and overtime. Their work emphasizes the impact of patient heterogeneity on optimal appointment schedules, suggesting segmenting schedules into overbooked and non-overbooked slots and optimizing patient order within these segments. Extending this, Zacharias and Yunes (2020) incorporate stochastic service times and demonstrate the multimodularity of the objective function.

Secondly, several overbooking heuristics have been proposed. Samorani and LaGanga (2015) address overbooking by considering individual appointment characteristics and day-dependent no-show predictions, introducing a heuristic that assigns same-day appointments to likely show-ups and future-day appointments to likely no-shows. Their heuristic shows improved performance compared to open access under different parameter settings. Additionally, Srinivas and Ravindran (2018) propose several scheduling heuristics based on patient-specific no-show risks to minimize waiting time, idle time, and overtime, while maximizing patient acceptance.

Thirdly, the importance of prediction accuracy has been demonstrated. Samorani and Harris (2019) employ probabilistic classifiers to predict individual no-show probabilities and optimize appointment schedules. They find that even slight improvements in prediction accuracy lead to significant reductions in schedule cost. Similarly, Li et al. (2019) develop a Bayesian nested logit model to enhance prediction accuracy and demonstrate reduced total cost incurred by waiting time, idle time, and overtime.

Although the proposed overbooking strategies are effective or even optimal in terms of cost minimization, they may raise ethical concerns, as shown by Samorani et al. (2021). Patients with a higher likelihood of no-shows may be disproportionately allocated undesirable appointment slots, leading to inequitable service experiences. Consequently, there is a need for appointment schedules that minimize costs while ensuring fairness. To address this, Samorani et al. (2021)

propose an objective function that strikes a balance between minimizing schedule costs and mitigating disparities in waiting times among different patient groups.

This work contributes to the overbooking literature by minimizing costs incurred by idle time and overtime while providing worst-case waiting time guarantees for all patients. It also offers a framework for considering no-show predictions of varying granularity and accuracy.

3.3. Problem Formulation

This section is organized as follows. In Section 3.3.1, we formally introduce the Robust Predictive Overbooking problem with Waiting Time Guarantees. In Section 3.3.2, we derive a mixed-integer linear program for this problem.

3.3.1. Mathematical Model

The notation and model builds upon that for the Robust Appointment Scheduling Problem with Waiting Time Guarantees (RASWTG) from Section 2.3, summarized in Appendix B.2.

Specifically, we consider a set of n preselected patients denoted by $[n] := \{1, \ldots, n\}$, and a single service provider with a regular working time of L time units. We aim to determine a patient sequence and appointment times that minimize the cost incurred by idle time and overtime, while ensuring that no patient $j \in [n]$ waits longer than their maximum waiting time guarantee W_j. Idle time before the first appointment incurs a per-time unit cost c_1, and idle time between two consecutive appointments $i-1$ and i incurs a per-time unit cost c_i, for $2 \leq i \leq n$. If there is any idle time before the planning horizon of L time units ends, it incurs a per-time unit cost c_{n+1}. Additionally, overtime incurs a per-time unit cost c_o. To determine the patient sequence, we introduce binary variables π_{ij} for $i, j \in [n]$, where $\pi_{ij} = 1$ if and only if exactly $i-1$ patients are served before patient j. In other words, patient j is assigned to the i-th appointment. We define continuous variables $A_i \geq 0$ to represent the start time of each appointment $i \in [n]$.

The main difficulty is that patients now not only have uncertain service times but also uncertain no-show behavior. Similar to Chapter 2, for each patient

$j \in [n]$, we only have information about the shortest and longest possible service times, denoted as $\underline{p}_j$ and $\bar{p}_j$, respectively. This leads to the scenario set of possible service time realizations $\mathcal{P} := [\underline{p}_1, \bar{p}_1] \times \ldots \times [\underline{p}_n, \bar{p}_n]$. In contrast to Chapter 2, we extend our analysis to include binary predictions for patients who do not show up. This involves considering a candidate set $\mathcal{C} \subseteq [n]$ of predicted no-shows, along with uncertainty budgets $\Gamma^{FP} \in \mathbb{N}_0$ for the total number of falsely predicted no-shows and $\Gamma^{FN} \in \mathbb{N}_0$ for the total number of falsely predicted show-ups. The confusion matrix derived from binary classification predictions of no-shows provides an evaluation of false positives and false negatives predictions based on historical data, serving as a reference for these uncertainty budgets. The scenario set of possible no-show realizations is then determined accordingly as

$$\Lambda := \left\{ \lambda \in \{0,1\}^n : \sum_{j \in \mathcal{C}} \lambda_j = \Gamma^{FP}, \sum_{j \in [n] \setminus \mathcal{C}} (1 - \lambda_j) = \Gamma^{FN} \right\} \tag{3.1}$$

where $\lambda_j = 1$ if and only if patient j shows up. The uncertainty budgets are enforced with equality, not inequality, to ensure the adversarial in the robust problem considers both false positives and false negatives, i.e., preventing either from being set to zero. It is noteworthy that the special case in which we estimate that a total of k patients show up, but lack individual predictions, is covered by setting $\mathcal{C} = \emptyset$ and $\Gamma^{FN} = n - k$ for some $k \leq n$.

For any given but fixed service time scenario $p \in \mathcal{P}$ and no-show scenario $\lambda \in \Lambda$, and given π and A, we compute waiting times, idle times, and overtime as follows. We start by initializing $C_0(\pi, A, p, \lambda) := 0$, and define completion times recursively for each appointment $i \in [n]$ as $C_i(\pi, A, p, \lambda) := \max\{A_i, C_{i-1}(\pi, A, p, \lambda)\} + \sum_{j=1}^{n} \pi_{ij} \lambda_j p_j$. It is useful and easy to see by induction that the completion times can be reformulated as $C_i(\pi, A, p, \lambda) = \max_{l \in [1,i]} \left\{ A_l + \sum_{k=l}^{i} \sum_{j=1}^{n} \pi_{kj} \lambda_j p_j \right\}$, see Lemma 3 in the Appendix. The waiting time before each appointment $i \in [n]$ then equals the duration between the completion of the previous appointment and the start of the current appointment, expressed as

$$\text{wait}_i(\pi, A, p, \lambda) := \sum_{j=1}^{n} \pi_{ij} \lambda_j \left(C_{i-1}(\pi, A, p, \lambda) - A_i \right)^+ \tag{3.2}$$

with $(\cdot)^+ := \max\{0, \cdot\}$. Similarly, the idle time before each appointment $i \in [n+1]$, with $A_{n+1} := L$, is given by $\text{idle}_i(\pi, A, p, \lambda) := (A_i - C_{i-1}(\pi, A, p, \lambda))^+$, and overtime is computed as $\text{over}(\pi, A, p, \lambda) := (C_n(\pi, A, p, \lambda) - L)^+$. The cost function

$$\begin{aligned}\text{cost}(\pi, A, p, \lambda) := &\sum_{i=1}^{n} c_i \cdot \text{idle}_i(\pi, A, p, \lambda) \\ &+ \max\left\{c_{n+1} \cdot \text{idle}_{n+1}(\pi, A, p, \lambda), c_o \cdot \text{over}(\pi, A, p, \lambda)\right\} \end{aligned} \tag{3.3}$$

accounts for the total cost arising from idle time and overtime.

Predictive overbooking should be performed sufficiently robust to withstand the adverse effects of unfavorable no-show occurrences, including unacceptable waiting times and increased costs. Therefore, we propose a robust optimization approach aimed at ensuring smooth operations for all stakeholders, even in worst-case scenarios. This gives rise to the Robust Predictive Overbooking problem with Waiting Time Guarantees (RPOWTG):

$$\min_{\pi, A} \max_{\substack{p \in \mathcal{P} \\ \lambda \in \Lambda}} \left\{\text{cost}(p, \lambda, \pi, A)\right\} \tag{3.4}$$

$$\text{s.t.} \max_{\substack{p \in \mathcal{P} \\ \lambda \in \Lambda}} \left\{ \max_{i \in [n]} \left\{ \text{wait}_i(p, \lambda, \pi, A) - \sum_{j=1}^{n} \pi_{ij} W_j \right\} \right\} \leq 0 \tag{3.5}$$

$$\sum_{i=1}^{n} \pi_{ij} = 1, \sum_{j=1}^{n} \pi_{ij} = 1 \qquad \forall i, j \in [n] \tag{3.6}$$

$$A_1 = 0 \tag{3.7}$$

$$\pi_{ij} \in \{0, 1\} \qquad \forall i, j \in [n] \tag{3.8}$$

RPOWTG extends the Robust Appointment Scheduling Problem with Waiting Time Guarantees by adding the consideration of no-show predictions. Objective Function (3.4) minimizes total worst-case cost incurred by idle time and overtime. Constraint (3.5) provides the well-known waiting time guarantees. Constraints (3.6) ensure that each patient is assigned to exactly one appointment, and vice versa. Constraint (3.7) sets the start time of the first appointment to 0.

Remark 5. The problem always has a feasible solution. For an arbitrary but fixed sequence π, it is easy to see that planning for all patients to show up with the appointment times $A_1 := 0$ and

$$A_i := \left(\max_{l \in [1,i-1]} \left(A_l + \sum_{k=l}^{i-1} \sum_{j=1}^{n} \pi_{kj} \bar{p}_j \right) - \sum_{j=1}^{n} \pi_{ij} W_j \right)^+ \tag{3.9}$$

for $i \in [2, n]$ are feasible solutions. In the special case that all waiting time guarantees are so long that the latter term equals 0 for all $i \in [n]$, all patients could be scheduled at time 0, which corresponds to the single-block rule in Cayirli and Veral (2003).

Remark 6. Due to the adversarial problems in Objective Function (3.4) and Constraint (3.5), which maximize the total cost and the waiting time of the longest-waiting patient, the above problem formulation is non-linear. In particular, it is not amenable to standard MILP solvers in this form.

Remark 7. The Robust Predictive Overbooking problem with Waiting Time Guarantees is $\mathcal{NP}$-hard, as it is as extension of RASWTG from Section 2.3.

3.3.2. Decomposition

We now derive and present the mixed-integer linear program for the Robust Predictive Overbooking problem with Waiting Time Guarantees. This involves decomposing the adversarial problems related to maximizing waiting times and the more difficult case of maximizing total cost.

3.3.2.1. Adversarial Waiting Time Maximization

We first show that the waiting time Constraint (3.5) can be decomposed by patients. Then we use total unimodularity and the classical dualization scheme by Bertsimas and Sim (2004) to linearize the individual waiting time constraints.

The following lemma establishes the equivalence of maximizing the waiting time for the longest waiting patient and independently computing a worst-case scenario for each patient that maximizes their waiting time. This inherent decomposability

of the adversary problem for waiting times stems from the scenario- independent appointment times.

Lemma 1 (Decomposability)**.** Given a patient sequence $(\pi_{ij})_{i,j\in[n]}$ and appointment times $(A_i)_{i\in[n]}$, Constraint (3.5) on waiting times is fulfilled if and only if

$$\max_{i\in[n]}\left\{\max_{p^{(i)}\in\mathcal{P},\lambda^{(i)}\in\Lambda}\left\{C_{i-1}(\pi,A,p^{(i)},\lambda^{(i)})-A_i-\sum_{j=1}^{n}\pi_{ij}W_j:\sum_{j=1}^{n}\pi_{ij}\lambda_j^{(i)}=1\right\}\right\}\leq 0.$$

Proof. Given a patient sequence π, appointment times A, and waiting time guarantees $W_1,\ldots,W_n$, it clearly holds

$$\max_{p\in\mathcal{P},\lambda\in\Lambda_{k-1}}\left\{\max_{i\in[n]}\left\{C_{i-1}(A,\pi,p,\lambda)-A_i-\sum_{j=1}^{n}\pi_{ij}W_j\right\}\right\} \tag{3.10}$$

$$\leq\max_{i\in[n]}\left\{\max_{p^{(i)}\in\mathcal{P},\lambda^{(i)}\in\Lambda_{k-1}}\left\{C_{i-1}(A,\pi,p^{(i)},\lambda^{(i)})-A_i-\sum_{j=1}^{n}\pi_{ij}W_j\right\}\right\}. \tag{3.11}$$

Assume the inequality holds strictly (i.e., with "$<$"). Let i^*,p^*,λ^* maximize the right-hand side. Then, for any $A\geq 0$, we obtain

$$C_{i^*-1}(A,\pi,p^*,\lambda^*)-A_{i^*}-\sum_{j=1}^{n}\pi_{i^*j}W_j \tag{3.12}$$

$$\leq\max_{i\in[n]}\left\{C_{i-1}(A,\pi,p^*,\lambda^*)-A_i-\sum_{j=1}^{n}\pi_{ij}W_j\right\} \tag{3.13}$$

$$\leq\max_{p\in\mathcal{P},\lambda\in\Lambda_{k-1}}\left\{\max_{i\in[n]}\left\{C_{i-1}(A,\pi,p,\lambda)-A_i-\sum_{j=1}^{n}\pi_{ij}W_j\right\}\right\} \tag{3.14}$$

$$<\max_{i\in[n]}\left\{\max_{p^{(i)}\in\mathcal{P},\lambda^{(i)}\in\Lambda_{k-1}}\left\{C_{i-1}(A,\pi,p^{(i)},\lambda^{(i)})-A_i-\sum_{j=1}^{n}\pi_{ij}W_j\right\}\right\} \tag{3.15}$$

$$=C_{i^*-1}(A,\pi,p^*,\lambda^*)-A_{i^*}-\sum_{j=1}^{n}\pi_{i^*j}W_j, \tag{3.16}$$

which is a contradiction. □

Thus, the problem of maximizing the waiting time for the longest-waiting patient

reduces to maximize, for each appointment, the previous appointment's completion time.

For each appointment $i \in [n]$, to maximize the previous appointment's completion time, we first use the non-recursive definition of the completion times to obtain $C_{i-1}(\pi, A, p, \lambda) = \max_{l \in [i-1]} A_l + \sum_{k=l}^{i-1} \sum_{j=1}^{n} \pi_{ij} \lambda_j p_j$. Now, as for each no-show scenario $\lambda \in \Lambda$, service time scenario $p \in \mathcal{P}$, and patient $j \in [n]$, the realized service time $\lambda_j p_j$ is non-negative across all scenarios, it is straightforward to see that all service times can be set to their maximum value $\bar{p}$. Therefore, the adversarial problem of maximizing waiting times reduces to finding a no-show scenario that maximizes

$$\max_{\lambda \in \Lambda} \left\{ \max_{l \in [1, i-1]} \left(A_l + \sum_{k=l}^{i-1} \sum_{j=1}^{n} \pi_{kj} \lambda_j \bar{p}_j) \right) : \sum_{j=1}^{n} \pi_{ij} \lambda_j = 1 \right\} \tag{3.17}$$

the completion time of the previous appointment. Now, similarly to the decomposability in Theorem 1, this is equivalent to maximize for each $l \in [i-1]$:

$$\max \sum_{k=l}^{i-1} \sum_{j=1}^{n} \pi_{kj} \bar{p}_j \lambda_j \tag{3.18}$$

$$\text{s.t.} \sum_{j \in \mathcal{C}} \lambda_j = \Gamma^{FP} - \sum_{j \in \mathcal{C}} \pi_{ij} \tag{3.19}$$

$$\sum_{j \in [n] \setminus \mathcal{C}} (1 - \lambda_j) = \Gamma^{FN} + \sum_{j \in [n] \setminus \mathcal{C}} \pi_{ij} \tag{3.20}$$

$$\lambda \in \{0, 1\}^n \tag{3.21}$$

Objective Function (3.18) maximizes the total service time between the lth and ith appointments. Constraint (3.19) ensures that exactly Γ^{FP} of the no-show candidates appear, while Constraint (3.20) ensures the correct number of false negatives to be considered. Specifically, it is implicitly ensured that the patient who is scheduled for the ith appointment also shows up, as otherwise, their waiting time is zero. Therefore, if a no-show candidate is scheduled for the ith appointment, at most $\Gamma^{FP} - 1$ of the no-show candidates may appear for other appointments. Similarly, if a show-up candidate is scheduled for the ith appointment, the adversarial will still choose the patient not to show up,

necessitating the distribution of Γ^{FN} no-shows besides the patient scheduled for the ith appointment.

The constraint matrix of (3.18) - (3.21) is totally unimodular, as can for example be seen by summing over all rows while multiplying the first row by -1. Let $\Gamma^{FN}_{\neg} := \Gamma^{FN} + |\mathcal{C}| - n$. Due to total unimodularity, the LP relaxation yields the same optimal solution as the original problem, and so does the LP dual of the LP relaxation:

$$\begin{aligned}
&\min\,(\Gamma^{FP} - \sum_{j\in\mathcal{C}} \pi_{ij}) z_{FP} \\
&\qquad + (\Gamma^{FN} + \sum_{j\in[n]\setminus\mathcal{C}} \pi_{ij} + |\mathcal{C}| - n) z_{FN} + \sum_{j=1}^{n} z_j && && (3.22)\\
&\text{s.t. } z_{FP} + z_j \geq \sum_{k=l}^{i-1} \pi_{kj}\bar{p}_j && \forall j \in \mathcal{C} && (3.23)\\
&\qquad - z_{FN} + z_j \geq \sum_{k=l}^{i-1} \pi_{kj}\bar{p}_j && \forall j \in [n]\setminus\mathcal{C} && (3.24)\\
&\qquad z_j \geq 0 && \forall j \in [n] && (3.25)\\
&\qquad z_{FP}, z_{FN} \in \mathbb{R} && && (3.26)
\end{aligned}$$

Therefore, the waiting time constraint in (3.5) can be reformulated, for all appointments $i \in [n]$ and $l \in [i-1]$, as

$$\begin{aligned}
&A_l + (\Gamma^{FP} - 1) z_{FP}^{(l,i)} + (\Gamma^{FN} + |\mathcal{C}| - n) z_{FN}^{(l,i)} + \sum_{j=1}^{n} z_j^{(l,i)} - A_i \\
&\qquad \leq \sum_{j=1}^{n} \pi_{ij} W_j + M \cdot \sum_{j\in[n]\setminus\mathcal{C}} \pi_{ij} && (3.27)\\
&A_l + \Gamma^{FP} z_{FP}^{(l,i)} + (\Gamma^{FN} + 1 + |\mathcal{C}| - n) z_{FN}^{(l,i)} + \sum_{j=1}^{n} z_j^{(l,i)} - A_i \\
&\qquad \leq \sum_{j=1}^{n} \pi_{ij} W_j + M \cdot \sum_{j\in\mathcal{C}} \pi_{ij}, && (3.28)
\end{aligned}$$

together with the Constraints (3.23) – (3.25) and a sufficiently large parameter M.

By employing the completion time reformulation, leveraging the total unimodularity of the constraint matrix associated with the adversarial problem for each appointment, and applying LP duality, the adversarial problem of maximzing waiting times has been reduced to $\mathcal{O}(n^2)$ linear programming problems. Each of these problems entails $\mathcal{O}(n)$ variables and constraints.

3.3.2.2. Adversarial Cost Maximization

The cost function is more difficult to linearize, as it is recursive and nonlinear in the no-show scenario λ. However, we discern structural properties by organizing costs into blocks. Initially, we partition costs into an idle time cost block followed by an overtime cost block, extending Theorem 2 to incorporate no-shows. Subsequently, we further subdivide the idle time cost into blocks that accumulate idle time. Leveraging this block structure enables us to formulate a mixed-integer linear program with a totally unimodular constraint matrix to find a cost-maximizing no-show scenario, akin to the waiting time adversarial's problem. However, this program only yields a maximizing no-show scenario rather than the maximum cost value. Thus, it must be integrated into a mixed-integer linear program for the Robust Predictive Overbooking problem with Waiting Time Guarantees along with the corresponding dual LP. Using strong duality, this integration ensures that the primal no-show scenario indeed maximizes total cost by enforcing the objective functions of the primal cost-maximizing problem and its corresponding LP dual to take an equivalent value.

Firstly, the following theorem establishes that the total cost is maximized under the same $n+1$ sequence-dependent service time scenarios as for RASWTG, where patients scheduled for the first $i^* \in \{0, \dots, n\}$ appointments require minimum service duration (maximizing idle time) and all subsequent patients require maximum service time (maximizing overtime). For simplicity of notation, let $C_i := C_i(\pi, A, p, \lambda)$. The proof is a straightforward extension of that of Theorem 2.

Theorem 5. *Given a patient sequence π, appointment times A, and $\lambda \in \Lambda$, let $\underline{C}_i(\pi, A, \lambda) := \max\{A_i, \underline{C}_{i-1}(\pi, A, \lambda)\} + \sum_{j=1}^{n} \pi_{ij}\lambda_j \underline{p}_j$ for $i \in [n]$, with $\underline{C}_0(\pi, A, \lambda) :=$*

0, *define the completion times when all patients require minimum service time. Then* $\max_{p\in\mathcal{P}} cost(p,\lambda,\pi,A)$ *is equivalent to*

$$\max_{i^*\in\{1,\dots,n+1\}} \sum_{i=1}^{i^*} c_i(A_i - \underline{C}_{i-1}(\pi,A,\lambda))^+ + c_o\left(A_{i^*} + \sum_{k=i^*}^{n}\sum_{j=1}^{n}\pi_{kj}\lambda_j\bar{p}_j - L\right)^+ \tag{3.29}$$

with $A_{n+1} := L$.

The proof is given in Appendix A.2.1. Now, we aim to characterize a worst-case no-show scenario for an arbitrary but fixed sequence and appointment times. Therefore, we first find a block structure of idle times:

Lemma 2. Let π be an assignment, $(A_i)_{i\in[n]}$ appointment times, $p\in\mathcal{P}$ a service time scenario, and $\lambda\in\Lambda$ a no-show scenario. Moreover, let $\{i\in[n] : (A_i - C_{i-1}) > 0\} = \{i_1,\dots,i_r\} =: I$ with $i_1 \le \dots \le i_r$ denote the index set of appointments before which positive idle time occurs. Then the idle time cost $\sum_{i=1}^{n+1} c_i(A_i - C_{i-1})^+$ is equivalent to

$$\sum_{l=1}^{r-1}(c_{i_l} - c_{i_{l+1}})A_{i_l} + c_{i_r}A_{i_r} - c_{i_1}A_1 - \sum_{l=1}^{r} c_{i_l}\sum_{\sigma=i_{l-1}}^{i_l-1}\lambda_\sigma p_\sigma \tag{3.30}$$

Proof. Assume w.l.o.g. that $\pi = id$, where id is the identity function. For each $i\in I$, let $l(i)$ denote the smallest index such that $C_{i-1} = A_{l(i)} + \sum_{\sigma=l(i)}^{i-1} p_\sigma\lambda_\sigma$ (see Lemma 3 in Appendix A.2.1).

We show that $l(i_j) = i_{j-1}$ for all $j\in[r]$, $j\ge 2$, via a proof by contradiction. First assume that $l(i_{j+1}) < i_j$, and for the purpose of simpler notation, let w.l.o.g. $j = 2$. On the one hand, Theorem 3 implies that $C_{i_2-1} \ge A_{i_1} + \sum_{\sigma=i_1}^{i_2-1} p_\sigma\lambda_\sigma$. On the other hand, it holds that $C_{i_2-1} = A_{l(i_2)} + \sum_{\sigma=l(i)}^{i_1-1} p_\sigma\lambda_\sigma + \sum_{\sigma=i_1}^{i_2-1} p_\sigma\lambda_\sigma \le C_{i_1-1} + \sum_{\sigma=i_1}^{i_2-1} p_\sigma\lambda_\sigma$. Together, it follows that $C_{i_1-1} \ge A_{i_1}$, a contradiction to $i_1\in I$. Therefore, $l(i_2)\ge i_1$.

Now assume that $l(i_2) > i_1$. On the one hand, as $l(i_2)$ is the smallest index with the above property, it must hold that $C_{i_2-1} > A_{i_1} + \sum_{\sigma=i_1}^{i_2-1} p_\sigma\lambda_\sigma$. On the other hand, because now $l(i_2)\notin I$, it holds that $A_i \le C_{i-1}$ for all $i\in[i_1+1, l(i_2)]$ and therefore $C_{i_2-1} = A_{l(i_2)} + \sum_{\sigma=l(i_2)}^{i_2-1} p_\sigma\lambda_\sigma \le C_{l(i_2)-1} + \sum_{\sigma=l(i_2)}^{i_2-1} p_\sigma\lambda_\sigma =$

$\max\{C_{l(i_2)-2}, A_{l(i_2)-1}\} + \sum_{\sigma=l(i_2)-1}^{i_2-1} p_\sigma \lambda_\sigma \leq \ldots \leq A_{i_1} + \sum_{\sigma=i_1}^{i_2-1} p_\sigma \lambda_\sigma$. Together this yields the contradiction $A_{i_1} \geq C_{i_2-1} - \sum_{\sigma=i_1}^{i_2-1} p_\sigma \lambda_\sigma > A_{i_1}$, from which it finally follows that $l(i_2) = i_1$.

The above argument for $l(i_2) \leq i_1$ can be equivalently applied to show that $l(i_1) = 1$, so we define $i_0 := 1$. Altogether, this yields:

$$\sum_{i=1}^{n+1} c_i \left(A_i - C_{i-1}\right)^+ \tag{3.31}$$

$$= \sum_{l=1}^{r} c_{i_l} \left(A_{i_l} - C_{i_l-1}\right) \tag{3.32}$$

$$= \sum_{l=1}^{r} c_{i_l} \left(A_{i_l} - A_{i_{l-1}} - \sum_{k=i_{l-1}}^{i_l-1} p_k \lambda_k \right) \tag{3.33}$$

$$= \sum_{l=1}^{r-1} \left(c_{i_l} - c_{i_{l+1}}\right) A_{i_l} + c_{i_r} A_{i_r} - c_{i_1} A_{i_0} - \sum_{l=1}^{r} c_{i_l} \sum_{\sigma=i_{l-1}}^{i_l-1} \lambda_\sigma p_\sigma \tag{3.34}$$

□

This block structure enables us to characterize a worst-case no-show scenario as follows:

Corollary 1. Let $i^* \in \{1, \ldots, n+1\}$. For any appointment $i \in [n]$, let $c(i) := c_{\min\{l \in [i,n]: A_l - \underline{C}_{l-1}(\pi, A, \lambda) > 0\}}$ denote the unit idle time cost of the next appointment before which a positive amount of idle time is incurred under minimum service times $\underline{p}$ and any no-show scenario $\lambda \in \Lambda$. If no idle time is incurred for any of these appointments, then we choose $c(i) := \max_{l \in [n]} c_l$.

Then $\lambda \in \Lambda$ maximizes the total cost

$$\max_{\lambda \in \Lambda} \left\{ \sum_{i=1}^{i^*} c_i (A_i - \underline{C}_{i-1}(\pi, A, \lambda))^+ + c_o \left(A_{i^*} + \sum_{k=i^*}^{n} \pi_{kj} \lambda_j \bar{p}_j - L \right)^+ \right\} \tag{3.35}$$

if and only if λ maximizes

$$\max_{\lambda \in \Lambda} \left\{ - \sum_{i=1}^{i^*-1} \sum_{j=1}^{n} \pi_{ij} c(i) \underline{p}_j \lambda_j^{(i^*)} + \sum_{i=i^*}^{n} \sum_{j=1}^{n} \pi_{ij} c_o \bar{p}_j \lambda_j^{(i^*)} \right\} \tag{3.36}$$

Proof. Following the case distinction in the proof of Theorem 5, it is sufficient to consider the cost $\sum_{i=1}^{i^*} c_i(A_i - \underline{C}_{i-1}(\pi, A, \lambda))^+ + c_o \left(A_{i^*} + \sum_{i=i^*}^{n} \pi_{ij}\lambda_j \bar{p}_j - L\right)^+$ for each $i^* \in \{0, \dots, n+1\}$ when maximzing total cost in the general case. To show the claim, we use the idle time reformulation from Lemma 2:

$$\sum_{i=1}^{i^*} c_i(A_i - \underline{C}_{i-1}(\pi, A, \lambda))^+ + c_o \left(A_{i^*} + \sum_{k=i^*}^{n} \pi_{kj}\lambda_j \bar{p}_j - L\right)^+ \tag{3.37}$$

$$= \sum_{l=1}^{r-1} (c_{i_l} - c_{i_{l+1}}) A_{i_l} + c_{i_r} A_{i_r} - c_{i_1} A_{i_0} - \sum_{l=1}^{r} c_{i_l} \sum_{\sigma=i_{l-1}}^{i_l - 1} \lambda_\sigma \underline{p}_\sigma \tag{3.38}$$

$$+ c_o \left(A_{i^*} + \sum_{k=i^*}^{n} \pi_{kj}\lambda_j \bar{p}_j - L\right)^+ \tag{3.39}$$

Given a sequence π and appointment times A, the last term is maximized by the no-show scenario described above. □

Thus, it suffices to consider, for each $i^* \in \{1, \dots, n+1\}$, a no-show scenario in which, among the patients scheduled for the last $n - i^* + 1$ appointments, those with the longest maximum service times show up. In case that more than $n - i^* + 1$ patients show up in total, then all patients scheduled for these last appointments show up. Additionally, among the patients scheduled for the first $i^* - 1$ appointments, the remaining ones with the smallest product of minimum service time and corresponding idle time cost show up. Notably, the idle time cost associated with customer j, who is scheduled for appointment i, is generally not the idle time cost c_i.

More formally, building on Corollary 1, a no-show scenario that maximizes cost for each $i^* \in [0, n+1]$ can be found by the following mixed-integer linear program:

$$\max -\sum_{i=1}^{i^*-1}\sum_{j=1}^{n}\pi_{ij}c(i)\underline{p}_j\lambda_j^{(i^*)} + \sum_{i=i^*}^{n}\sum_{j=1}^{n}\pi_{ij}c_o\bar{p}_j\lambda_j^{(i^*)} \tag{3.40}$$

$$\text{s.t.} \sum_{j\in\mathcal{C}}\lambda_j^{(i^*)} = \Gamma^{FP} \tag{3.41}$$

$$\sum_{j\in[n]\setminus\mathcal{C}}(1-\lambda_j^{(i^*)}) = \Gamma^{FN} \tag{3.42}$$

$$\lambda^{(i^*)} \in \{0,1\}^n \tag{3.43}$$

The constraint matrix of (3.40) – (3.43) is similar to that of (3.18) – (3.21) and therefore totally unimodular. Hence, strong duality holds for its LP relaxation, meaning that if an optimal solution of (3.40) – (3.43) exists, its value equals that of the following LP:

$$\min \Gamma^{FP}v_{i^*}^{FP} + (\Gamma^{FN} + |\mathcal{C}| - n)v_{i^*}^{FN} + \sum_{j=1}^{n}\underline{z}_{i^*j} \tag{3.44}$$

$$\text{s.t. } v_{i^*}^{FP} + \underline{z}_{i^*j} \geq \sum_{i=1}^{i^*-1}\pi_{ij}c(i)\underline{p}_j \qquad \forall j\in\mathcal{C} \tag{3.45}$$

$$-v_{i^*}^{FN} + \underline{z}_{i^*j} \geq \sum_{i=1}^{i^*-1}\pi_{ij}c_o\bar{p}_j \qquad \forall j\in[n]\setminus\mathcal{C} \tag{3.46}$$

$$\underline{z}_{i^*j} \geq 0 \qquad \forall j\in[n] \tag{3.47}$$

Now, aiming for a worst-case no-show scenario that maximizes total cost, we can integrate it into a mixed-integer linear program by incorporating the primal constraints (3.41) – (3.43) and the dual constraints (3.45) – (3.47), while ensuring that the objective function value (3.40) equals that of the dual (3.44).

Additionally, except for well-known linearizations such as the product of a binary and a non-negative continuous variable, it only remains to linearize the idle time cost function $c(i)$, as shows in (A.20) – (A.23) in the Appendix.

Remark 8 (Uncertainty Budgets for Service Times)**.** When additionally introducing uncertainty budgets for service times, it seems that the linearization can

be performed following a similar framework, with the additional linearization of the product of maximum service time deviation and no-show scenario.

3.3.3. Mixed-Integer Linear Program

The linearization results from Section 3.3.2 lead to a comprehensive MILP for the Robust Predictive Overbooking Problem with Waiting Time Guarantees. This MILP formulation uses binary variables λ_{i^*ij} for each case $i \in [n+1]$ from Theorem 5 to indicate not only, as in the sequence π, whether patient j is scheduled for the i-th appointment, but also whether patient j shows up. Additionally, let $U \geq 0$ be an upper bound on the worst-case total costs.

$$\min U \tag{3.48}$$

$$\text{s.t. } U \geq \sum_{i=2}^{i^*} c_i \Delta_{i^*i} + c_o \sigma_{i^*} \qquad \forall i^* \in [n+1] \tag{3.49}$$

$$\sigma_{i^*} \geq A_{i^*} + \sum_{k=i^*}^{n} \sum_{j=1}^{n} \lambda_{i^*ij} \bar{p}_j - L \qquad \forall i^* \in [n+1] \tag{3.50}$$

$$\Delta_{i^*i} \geq A_i - \underline{C}_{i^*\,i-1} \qquad \forall i^*, i \in [n+1] \tag{3.51}$$

$$\underline{C}_{i^*i} = \max\{A_i, \underline{C}_{i^*i-1}\} + \sum_{j=1}^{n} \lambda_{i^*ij} \underline{p}_j \qquad \forall i \in [n], i^* \in [n+1] \tag{3.52}$$

$$\sum_{i=1}^{n} \sum_{j \in \mathcal{C}} \lambda_{i^*ij} = \Gamma^{FP} \qquad \forall i^* \in [n+1] \tag{3.53}$$

$$\sum_{i=1}^{n} \sum_{j \in [n] \setminus \mathcal{C}} (1 - \lambda_{i^*ij}) = \Gamma^{FN} \qquad \forall i^* \in [n+1] \tag{3.54}$$

$$\begin{aligned} &- \textstyle\sum_{i=1}^{i^*-1} \sum_{j=1}^{n} c(i^*, i) \underline{p}_j \lambda_{i^*ij} + \sum_{i=i^*}^{n} \sum_{j=1}^{n} c_o \bar{p}_j \lambda_{i^*ij} \\ &\quad = \Gamma^{FP} v_{i^*}^{FP} + \Gamma_{\neg}^{FN} v_{i^*}^{FN} + \sum_{j=1}^{n} \underline{z}_{i^*j} \qquad \forall i^* \in [n+1] \end{aligned} \tag{3.55}$$

$$v_{i^*}^{FP} + \underline{z}_{i^*j} \geq \sum_{i=1}^{i^*-1} \pi_{ij} c(i^*, i) \underline{p}_j \qquad \forall j \in \mathcal{C}, i^* \in [n+1] \quad (3.56)$$

$$- v_{i^*}^{FN} + \underline{z}_{i^*j} \geq \sum_{i=1}^{i^*-1} \pi_{ij} c_o \bar{p}_j \qquad \forall j \in [n] \setminus \mathcal{C}, i^* \in [n+1] \quad (3.57)$$

$$c(i^*, i) = \sum_{l=i}^{n+1} c_l \gamma_{i^* il} \text{ with (A.20) – (A.23)} \qquad \forall i \in [n], i^* \in [n+1] \quad (3.58)$$

$$\mathrm{A}_l - A_i + (\Gamma^{FP} - \textstyle\sum_{j \in \mathcal{C}} \pi_{ij}) w_{li}^{FP} + (\Gamma^{FN} + \sum_{j \in [n] \setminus \mathcal{C}} \pi_{ij} + |\mathcal{C}| - n) w_{li}^{FN}$$
$$+ \sum_{j=1}^{n} \bar{z}_{lij} \leq \sum_{j=1}^{n} \pi_{ij} W_j \qquad \forall i \in [n], l \in [i, n] \quad (3.59)$$

$$w_{li}^{FP} + \bar{z}_{lij} \geq \sum_{k=l}^{i-1} \sum_{j=1}^{n} \pi_{kj} \bar{p}_j \qquad \forall j \in \mathcal{C}, i \in [n], l \in [i, n] \quad (3.60)$$

$$- w_{li}^{FN} + \bar{z}_{lij} \geq \sum_{k=l}^{i-1} \sum_{j=1}^{n} \pi_{kj} \bar{p}_j \qquad \forall j \in [n] \setminus \mathcal{C}, i \in [n], l \in [i, n] \quad (3.61)$$

$$\sum_{j=1}^{n} \pi_{ij} = 1, \sum_{i=1}^{n} \pi_{ij} = 1, \lambda_{i^* ij} \leq \pi_{ij} \qquad \forall i, j \in [n], i^* \in [n+1] \quad (3.62)$$

$$\underline{C}_{i^*0} = 0, A_{n+1} = L \qquad \forall i^* \in [n+1] \quad (3.63)$$

$$A_i, \Delta_{i^* i}, \sigma_{i^*} \geq 0 \qquad \forall i \in [n], i^* \in [n+1] \quad (3.64)$$

$$\underline{z}_{i^* j}, \bar{z}_{lij} \geq 0 \qquad \forall l, i, j \in [n], i^* \in [n+1] \quad (3.65)$$

$$\pi_{ij}, \lambda_{i^* ij} \in \{0, 1\} \qquad \forall i, j \in [n], i^* \in [n+1] \quad (3.66)$$

$$\gamma_{i^* il} \in \{0, 1\} \qquad \forall i, l \in [n], i^* \in [n+1] \quad (3.67)$$

Objective Function (3.48) minimizes the upper bound U on the total cost, computed through Constraints (3.49) – (3.57). Constraint (3.49) establishes the upper limit for costs arising from idle time $\Delta_{i^* i}$ and overtime σ_{i^*} in each worst-case scenario indexed by $i^* \in [n+1]$, as detailed in Theorem 5. Specifically, Constraint (3.50) defines the worst-case overtime accumulated starting from the i^*th appointment. Constraints (3.51) – (3.57) define the worst-case idle time before the ith appointment as a function of λ: The idle time before an appointment is the maximum of zero and the difference between the scheduled appointment time and the completion time of the previous appointment, as specified by Constraint (3.51). The completion times are determined by Constraint (3.52).

Constraints (3.53) – (3.54) ensure that the predicted number of patients show up, adjusted by a cost-maximizing set of patients based on the uncertainty budgets Γ^{FP} and Γ^{FN}. Building on Corollary 1 and duality, Constraints (3.55) – (3.57) ensure that the objective function value of the cost adversarial's problem (3.40) – (3.43) equals, for each case $i^* \in [n+1]$, that of the dual LP (A.39) – (3.47). In other words, they ensure that, for each case $i^* \in [n+1]$, as many patients for the last appointments show up as possible, and if not all can show up, then those with the longest maximum service times show up. Among the first i^* appointments, only those with the smallest product $c(i)\underline{p}_j$ show up. Finally, Constraint (3.58) computes the idle time cost $c(i)$ associated with appointment $i \in [n]$. Constraint (3.63) ensures a feasible assignment of patients to appointments. Constraints (3.59) – (3.61) enforce the waiting time guarantees derived in Section 3.3.2.1.

The program (3.48) – (3.67), when replacing Constraint (3.59) on the waiting times with the big-M formulation known from inequalities (3.27) – (3.28) and using standard linearization techniques for the maximum term in Constraint (3.52) as well as for the product of $c(i^*, i)\lambda_{i^*ij}$ in Constraints (3.56), (3.57), and (3.59), transforms the problem into a mixed-integer linear program. This MILP encompasses $O(n^3)$ binary variables, $O(n^4)$ continuous variables, as well as $O(n^3)$ constraints. The computational study in Section 3.4 demonstrates the efficient solvability for all instances involving up to eight patients from a data set of CT examinations in a large hospital, employing a state-of-the-art MILP solver.

Remark 9 (Constant Idle Time Costs)**.** In the case of constant idle time costs, Constraint (3.59) becomes redundant, as $c(i^*, i) = c_i$. Consequently, the mixed-integer linear program becomes significantly more efficient.

Remark 10 (Scheduling and sequencing rules)**.** The As-Soon-As-Possible scheduling rule and the Smallest Variance First sequencing rule with waiting time guarantees from Section 2.4 still yield optimal solutions under the assumption of zero no-shows. Furthermore, extending the ASAP scheduling rule to account for no-shows is straightforward, replacing the service time term in Line 3 with the optimal value of the adversarial's waiting time dual problem (3.22) to (3.25). Thus, given non-increasing idle time costs and any sequence π, it remains optimal to schedule all appointments as early as possible.

3.4. Computational Study

In this section, we assess the performance of RPOWTG schedules in a simulation based on real-world data from the radiology department of a large university hospital. Radiology departments are particularly affected by the negative effects of no-shows due to the significant capital investment in imaging equipment and the expertise required for their operation and evaluation. Our evaluation includes an analysis of runtimes, a sensitivity analysis, and managerial insights derived from comparing waiting times, idle times, and overtime under RPOWTG schedules to those obtained when disregarding no-shows, as discussed in Chapter 2.

3.4.1. Instances

The test instances are generated based on data collected from one of three CT machines in the radiology department of a large university hospital in Munich, Germany. For benchmarking purposes, the test set comprises the identical 1,386 CT examinations conducted in the first quarter of 2020 as discussed in Chapter 2. Each examination record contains details such as exam type, exam day, start time, and completion time. The only available information regarding no-shows is the estimation provided by the radiology department, indicating that daily no-show rates are not expected to surpass 20%.

To create complete test instances, for each examination, we utilize the service time interval estimated in Section 2.5. Additionally, we fix the following parameters consistently with the base case from Section 2.5: waiting time guarantees of 30 minutes for all patients (Hill and Joonas 2005), constant idle time costs of 1 unit per minute of idle time, and overtime costs of 1.25 units per minute. Subsequently, we vary patient count and no-show predictions using a full factorial design.

Number of Patients. To assess runtimes, we vary the patient counts to 5, 8, and 10 patients to be scheduled, selecting the respective number of first patients reported on each day that contains at least the specified number of examination reports. This results in 252 instances across all classes with five patients, 132 instances across all classes with eight patients, and 107 instances across all classes

with ten patients. Due to runtime constraints, we conduct the sensitivity analysis based on eight patients.

Uncertainty of No-Show Predictions. We simulate various levels of no-show predictions, ranging from no-show rates of 0% to 20%, and assess their impact on scheduling outcomes. These levels represent different scenarios of prediction uncertainty: perfect predictions, which are 100% accurate with no false positives or negatives; zeroFP predictions, where all predicted no-shows indeed do not show up, but some false negative predictions actually show up; zeroFN predictions, where all predicted show-ups show up, but some predicted no-shows are actually false positives; mixed predictions, involving both false positives and negatives; and no predictions, considering only the total no-show rate without specific candidate predictions and thus the total numbers of now shows for a given number of n patients. These scenarios allow us to evaluate the robustness of appointment schedules against the uncertainty inherent in no-show predictions.

The planning horizon is determined by multiplying the sum of the midpoints of the service time intervals by the anticipated show-up rate. For instances with eight patients, this results in an average planning horizon of 83.83 minutes in the case of 0% no-shows, 73.35 minutes for 10% no-shows, and 62.87 minutes for 20% no-shows. The average cumulative service time uncertainty within this horizon, calculated as the corresponding sum of service time interval lengths, is 78.65 minutes for 0% no-shows, 68.82 minutes for 10% no-shows, and 58.98 minutes for

Table 3.1. Levels of no-show predictions

Level	No-show rate	Candidates	Γ^{FP}	Γ^{FN}
0%_no_pred	0%	$\emptyset$	-	-
10%_no_pred	10%	$\emptyset$	-	$\lfloor 0.1 \cdot n \rfloor$
10%_mixed_pred	10%	$\{1, \ldots, \lfloor 0.2 \cdot n \rfloor\}$	$\lfloor 0.1 \cdot n \rfloor$	
10%_perfect_pred	10%	$\{1, \ldots, \lfloor 0.1 \cdot n \rfloor\}$	-	-
20%_no_pred	20%	$\emptyset$	-	$\lfloor 0.2 \cdot n \rfloor$
20%_mixed_pred	20%	$\{1, \ldots, \lfloor 0.2 \cdot n \rfloor\}$	$\lfloor 0.1 \cdot n \rfloor$	$\lfloor 0.1 \cdot n \rfloor$
20%_zeroFN_pred	20%	$\{1, \ldots, \lfloor 0.3 \cdot n \rfloor\}$	$\lfloor 0.1 \cdot n \rfloor$	-
20%_zeroFP_pred	20%	$\{1, \ldots, \lfloor 0.1 \cdot n \rfloor\}$	-	$\lfloor 0.1 \cdot n \rfloor$
20%_perfect_pred	20%	$\{1, \ldots, \lfloor 0.2 \cdot n \rfloor\}$	-	-

20% no-shows.

To evaluate schedules, we use the service times calculated as the difference between reported completion time and start time. In all scenarios, the same candidates as predicted in the corresponding scenario of perfect predictions show up. This allows us to assess the costs associated with false positive and false negative predictions.

3.4.2. Runtimes

In Table 3.2, we report the average runtimes to compute an optimal schedule using the mixed-integer program from Section 3.3.3. All experiments have been performed on an Intel Core i7-10510U 1.8 GHz machine (in 64-bit mode) with 16 GB memory under Microsoft Windows 10 and Gurobi 9.5.0 in Python 3.9.0.

First of all, runtimes increase substantially with the number of patients to be scheduled. Instances involving up to 8 patients can typically be solved to optimality within seconds or minutes. However, not all instances with 10 patients can be solved to optimality within the time limit of 3,600 seconds, especially when predictions entail higher uncertainty. This underscores the need for future research to focus on enhancing solution efficiency.

Moreover, runtimes increase with the extent of uncertainty under consideration.

Table 3.2. Average runtimes to compute an optimal schedule
(in seconds, percentage of instances solved to optimality within one hour for 10 patients)

	Patients			
No-show prediction	5	8		10
0%_no_pred	1.17	3.94	32.37	(100.00%)
10%_no_pred	21.57	140.46	711.17	(50.00%)
10%_mixed_pred	1.46	19.05	305.91	(82.61%)
10%_perfect_pred	0.65	6.69	255.67	(93.23%)
20%_no_pred	17.82	307.32	1,365.33	(12.12%)
20%_mixed_pred	5.11	184.99	1426.90	(9.09%)
20%_zeroFN_pred	3.24	25.47	237.77	(72.00%)
20%_zeroFP_pred	1.93	250.81	170.76	(42.43%)
20%_perfect_pred	1.28	5.82	309.57	(84.83%)

Specifically, they increase across scenarios of perfect, zeroFN, zeroFP, mixed, to no predictions, alongside the rise in the no-show rate from 0%, 10%, to 20%. We attribute this to the varying number of no-show scenarios under consideration: With perfect predictions there is only one scenario, namely that all patients show up. With zeroFN, no-shows are confined to the predicted set, while with zeroFP, they occur among the predicted show-ups. Given the smaller candidate set of no-shows compared to show-ups in our test instances, runtimes are notably longer for zeroFP compared to zeroFN predictions. Mixed predictions encompass both zeroFP and zeroFN predictions separately. With no predictions, schedules need to minimize costs and ensure waiting time guarantees, regardless of which Γ^{FN} of all patients do not show up.

3.4.3. Sensitivity Analysis

We now examine the base case of scheduling eight patients, each with a maximum waiting time guarantee of 30 minutes. Specifically, we conduct a sensitivity analysis by comparing the impact of different scenarios of no-show predictions on waiting times, idle times, and overtime, as summarized in Table 3.3. Notably, patients who do not show up are excluded when calculating waiting time. Moreover, the slight deviation of 100% in the fulfillment of waiting time guarantees is again due to the estimation of minimum and maximum service times derived from empirical service time data in the separate interval estimation set, excluding outliers (see Section 2.5.1). In cases where realized service times are outside the estimated range, the waiting time guarantees of a few later patients are unmet. Finally, we derive practical managerial insights.

It is important to recall several trade-offs with regard to capacity utilization. First of all, we measure capacity utilization through idle times, with increasing utilization entailing decreasing idle times. However, decreasing idle times generally leads to longer waiting times, and vice versa. Furthermore, while reducing idle time before any appointment is generally also advantageous for minimizing overtime, there is a direct trade-off between overtime and idle time after the last appointment.

3.4.3.1. Waiting Time

A summary of waiting times is provided in Table 3.3, complemented by a boxplot of waiting times for all levels of no-show predictions in Figure 3.1. Most importantly, the 30-minute waiting time guarantee is fulfilled for more than 96% of patients across all prediction levels. Notably, adherence improves as no-show rates increase due to heightened caution, and achieving 100% adherence is practically unattainable due to extreme cases. Moreover, the boxplot illustrates that while

Table 3.3. Average results by level of prediction uncertainty (in minutes, WTG: waiting time guarantee)

	Total idle time	Average wait	WTG fulfilled	Overtime
0%_no_pred:	16.73	11.49	96%	4.58
10%_no_pred:	14.78	8.69	98%	4.54
10%_mixed_pred:	14.30	10.80	96%	4.06
10%_perfect_pred:	13.80	11.55	96%	3.56
20%_no_pred:	13.23	7.15	98%	4.13
20%_mixed_pred:	12.40	8.19	96%	3.30
20%_zeroFN_pred:	12.63	10.35	98%	3.53
20%_zeroFP_pred:	12.39	8.72	98%	3.29
20%_perfect_pred:	12.24	11.16	97%	3.14

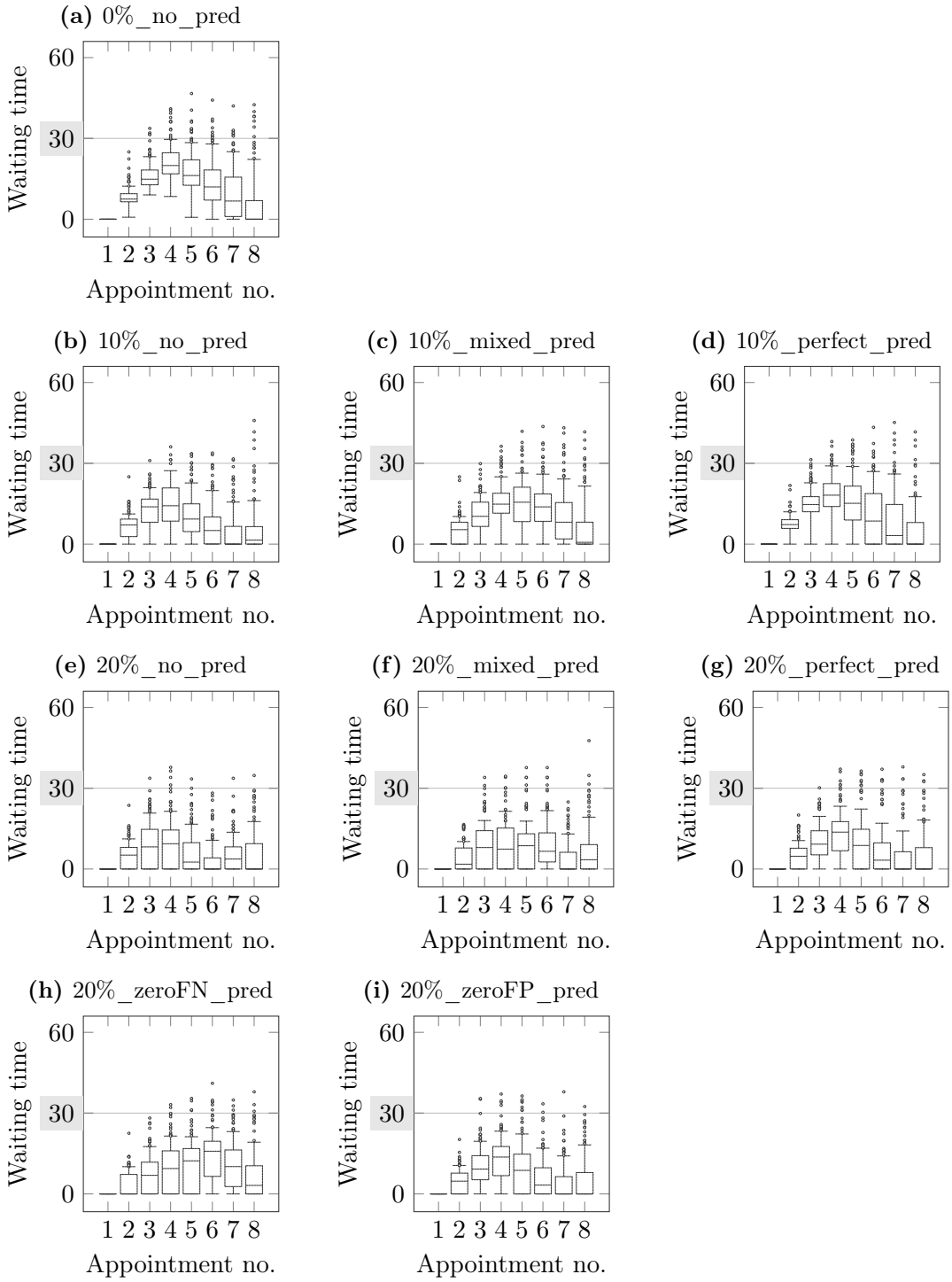

Figure 3.1. Waiting times by no-show rate and prediction uncertainty (in minutes)

some patient waiting times approach and a few exceed the 30-minute guarantee, most patients experience considerably shorter waiting times.

Average waiting times for attended appointments decrease with higher no-show rates, for the case of no predictions from 11.49 minutes with 0% no-shows to 7.15 minutes (-37.78%) with 20% no-shows, reflecting increased robustness against no-shows. Furthermore, higher prediction certainty leads to further reductions in average waiting times. For 20% no-shows, this results in a decrease from 11.16 minutes with perfect predictions to 7.15 minutes (-35.93%) with no predictions, as more accurate planning enables enhanced capacity utilization while still meeting waiting time guarantees. Interestingly, the case of zeroFP exhibits lower average waiting times than zeroFN, and waiting times show a more equitable distribution with higher predicted no-show rates, as further discussed in Section 3.4.3.4.

3.4.3.2. Overtime

A summary of overtime is presented in Table 3.3, accompanied by a boxplot illustrating overtime per no-show rate and level of prediction uncertainty in Figure 3.2. In over 87% of cases, overtime is less than 10 minutes. Overtime tends to decrease as prediction accuracy improves, as less buffer time is needed to ensure adherence to the waiting time guarantees. For instance, in the case of 20% no-shows, average overtime decreases by 23.97% from 4.13 minutes to 3.14 minutes when comparing no predictions to perfect predictions. Furthermore, in scenarios with no predictions, overtime decreases from 4.58 minutes with 0% no-shows to 4.13 minutes (-9.83%) with 20% no-shows.

3.4.3.3. Idle Time

A summary of the idle times is provided in Table 3.3. Figure 3.3 presents boxplots for idle time by appointment across all levels of no-show predictions. The idle time before each appointment, which conflicts with waiting times, is drawn in black, while the idle time after the last appointment, conflicting with overtime, is drawn in gray. Notably, there is almost no idle time before any of the first five appointments, and in some cases even before any appointment. We also note that idle time before later appointments could be further reduced with the introduction of uncertainty budgets for service times, as discussed in Section 2.5.3.2.

Moreover, the more accurate the predictions, the lower the total idle time. In scenarios with 10% no-shows, the total sum of idle time decreases from 14.78 minutes to 13.80 minutes (-6.63%) when transitioning from no predictions to perfect predictions. Similarly, in scenarios with 20% no-shows, the total sum of idle time decreases by 7.48% with perfect predictions compared to no predictions. Interestingly, zeroFP predictions allow for slightly higher capacity utilization than zeroFN predictions (12.39 minutes idle time compared to 12.63 minutes).

Surprisingly, as the level of no-shows increases, not only do waiting times for attendees and overtime decrease, but total idle time also decreases: In the case of no predictions with 0% no-shows, it is 16.73 minutes, with 10% no-shows, it is 14.78 minutes, and with 20% no-shows, it is 13.23 minutes. Even when breaking down idle time per number of show-ups (plus 1 for idle time after the last appointment), idle time remains remarkably stable: For scenarios with no predictions and 0% no-shows, the average idle time before each of the eight show-ups (plus one) is 1.86 minutes. With 10% no-shows, the average idle time before each of the seven show-ups (plus one) is 1.85 minutes. In scenarios with 20% no-shows, the average idle time before each of the six show-ups (plus one) is 1.89 minutes. This suggests that the proposed overbooking approach is capable of effectively mitigating negative effects implied by no-shows.

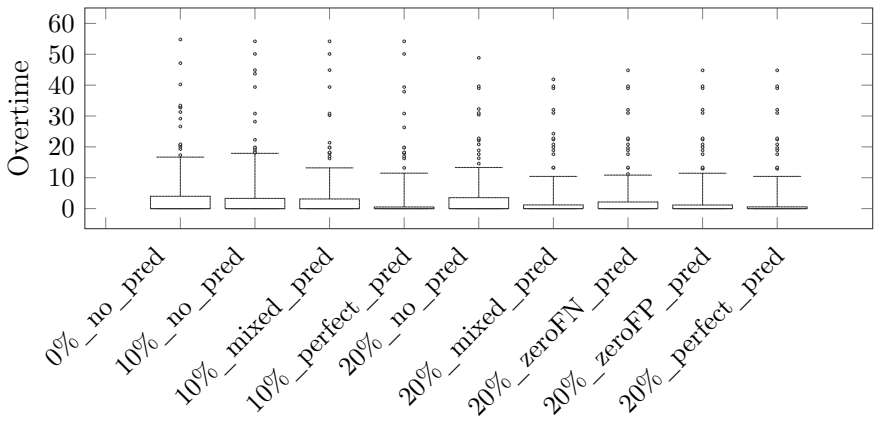

Figure 3.2. Overtime by no-show rate and prediction uncertainty (in minutes)

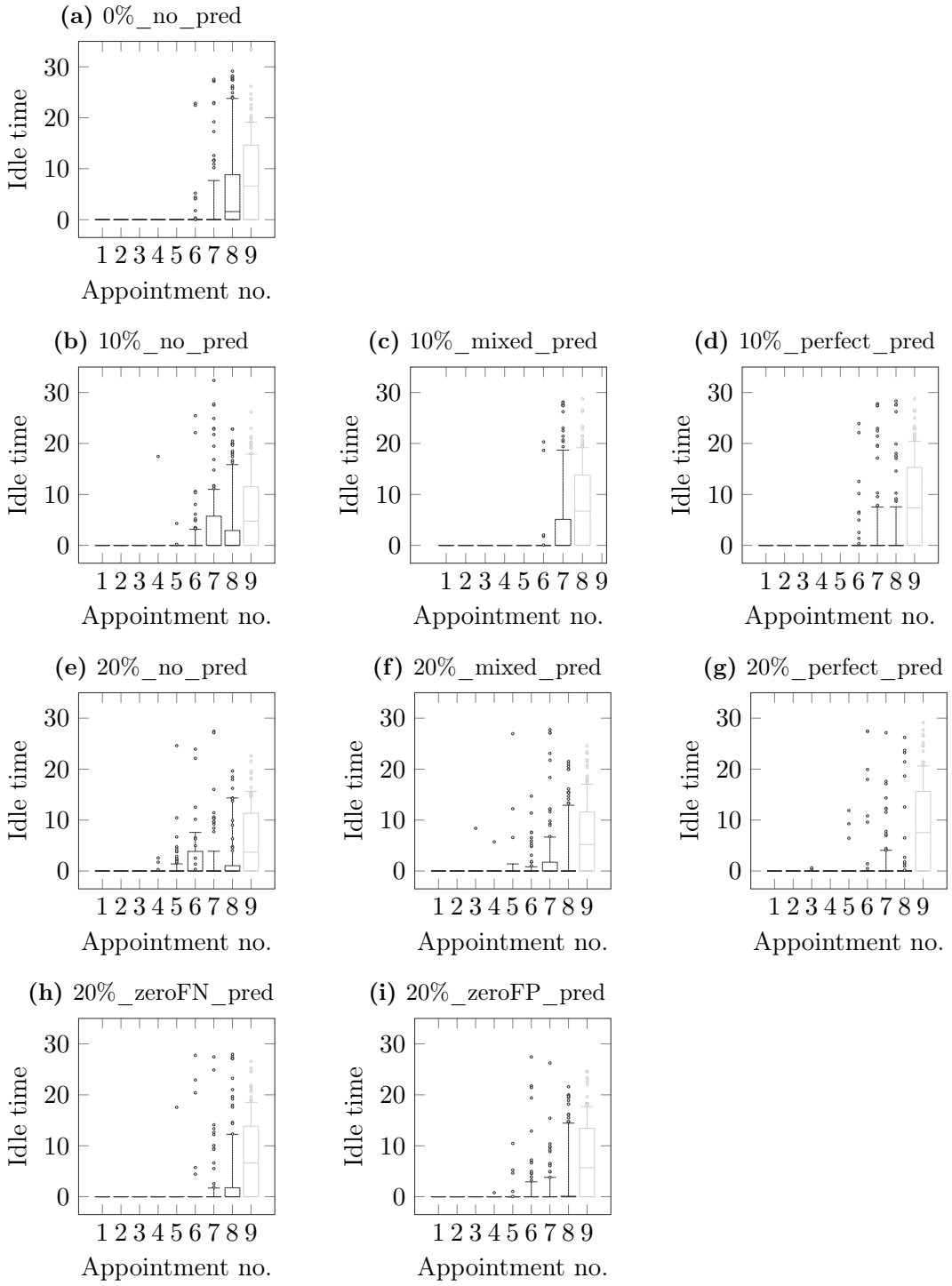

Figure 3.3. Idle times by no-show rate and prediction uncertainty (in minutes)

Finally, it is noteworthy that underestimating the no-show rate is preferable to not considering no-shows at all, while overestimating the rate must be avoided to maintain waiting time guarantees.

3.4.3.4. Discussion

Our findings align with and even surpass the expectations set by Samorani et al. (2021), indicating the viability of near-optimal overbooking schedules that achieve similar outcomes to traditional objective functions while being more equitable. Specifically, together with the comparison to a weighted sum method in Section 2.5.3.2, they illustrate that predictive overbooking with waiting time guarantees can increase capacity utilization, reduce overtime, and ensure reliable waiting time guarantees.

Interestingly, zeroFP predictions outperform zeroFN predictions across all criteria. This may seem counterintuitive considering the actual candidate set to consider for the no-shows, which is $\mathcal{C}$ for zeroFN predictions and the larger set $[n] \setminus \mathcal{C}$ for zeroFP predictions. However, this may potentially be attributed to the ratio of no-shows to the respective candidate sets. With zeroFN predictions, the ratio is $\frac{|\mathcal{C}|-\Gamma^{FP}}{|\mathcal{C}|}$, while with zeroFP predictions, it is $\frac{\Gamma^{FN}}{n-|\mathcal{C}|}$. Further investigation is needed to explore the impact of this ratio on optimal appointment schedules, as well as to understand this effect.

Surprisingly, our overbooking approach demonstrates that a higher no-show rate leads to improvements in waiting times, idle times, and overtime. This unexpected finding warrants further research, including exploring the interplay with service time uncertainty, as well as examining the positioning of predicted and potential no-show candidates in appointment schedules and their implications on appointment times.

3.5. Conclusion

This work presents a robust optimization approach to predictive overbooking. In the framework provided, we propose to use the uncertainty budgets introduced by Bertsimas and Sim (2004) to account for false positive and false negative predictions. Secondly, we introduce waiting time guarantees for overbooking,

which ensure that even in a worst-case scenario of no-show and service time realizations, certain waiting time thresholds are not exceeded for almost any patient. Under these hard constraints, the total cost incurred by idle time and overtime is minimized.

The Robust Predictive Overbooking Problem with Waiting Time Guarantees is shown to be decomposable into multiple layers. Initially, the worst-case waiting time constraint is broken down on a per-patient basis, allowing for its linearization through total unimodularity and LP duality. Secondly, the objective function is decomposed into two distinct components: an earlier idle time segment and a subsequent overtime incurring segment. Additionally, the idle time is further divided into blocks. Leveraging these decompositions, along with once more total unimodularity and LP duality, enables the linearization of the cost function. This results in a mixed-integer linear program.

Our computational results show that a state-of-the-art MILP solver can find optimal solutions in reasonable runtime for instances with up to eight patients. The findings demonstrate consistent adherence to the waiting time guarantees in over 96% of cases across all scenarios. Total idle time decreases by up to 8% with prediction accuracy, albeit with the counterintuitive observation that false positives seem to have a slightly more adverse effect than false negatives. Interestingly, a higher no-show rate decreases waiting times and simultaneously enables a cost reduction through our overbooking approach. This finding not only underscores that overbooking is a beneficial strategy for addressing the negative effects of no-shows. While requiring further validation in other settings, it also implies that targeting the direct mitigation of no-shows, such as through reminder systems - which may be beneficial in certain contexts, particularly when prediction accuracy can be improved - warrants careful cost-benefit assessment.

Finally, we suggest several avenues for future research. Firstly, exploring optimal scheduling and sequencing rules for robust predictive overbooking with waiting time guarantees, or other efficient solution methods for larger instance sizes, seems valuable. Additionally, introducing uncertainty budgets for service times, akin to the approach with respect to false no-show predictions, could offer further control over the level of conservatism. The problem setting could further be extended to integrate decision-making on accepting or rejecting patients. Secondly, our work illustrates the impact of false positives and false negatives

on overbooking. Closing the loop from prescriptive to predictive analysis, this warrants further analysis and possible adaptation of prediction methods. Lastly, the concept of employing uncertainty budgets for false predictions is promising for further investigation with respect to binary predictions and beyond, as well as for application to various other problems.

4. Robust Routing and Scheduling of Home Healthcare Workers

Co-authors: Jonathan Bard and Rainer Kolisch

Working paper (*Submitted* at *Transportation Science* at the time of submission of this dissertation).

Abstract. The global home healthcare market is growing rapidly due to aging populations, advancements in healthcare technology, and patient preference for home-based care. In this work, we study the multi-day planning problem of simultaneously deciding patient acceptance, assignment, routing, and scheduling under uncertain travel and service times. Our approach ensures cardinality-constrained robustness with respect to timely patient care and the prevention of overtime. We take into account a wide range of criteria including patient time windows, caregiver availability and compatibility, a minimum time interval between two visits of a patient, the total number of required visits, continuity of care, and profit. We use a novel systematic modeling scheme that prioritizes health-related criteria as hard constraints and optimizes cost and preference-related criteria as part of the objective function. We present a mixed-integer quadratic linear formulation, along with a nested branch-and-price technique. Results from a case study in Austin, Texas demonstrate that instances of realistic size can be solved to optimality within reasonable runtimes. The price of robustness is mainly due to reduced patient load per caregiver. Interestingly, the criterion of geographical proximity appears to be of secondary priority when selecting new patients and assigning them to caregivers.

4.1. Introduction

Home healthcare accounted for 3.5% of national health expenditures in the United States in 2019 (Centers for Medicare & Medicaid Services 2020). The global home healthcare market, valued at USD 336.0 billion in 2021, is predicted to expand at a compound annual growth rate of 7.9% from 2022 to 2030 (Grand View Research 2021). This substantial growth is attributed to factors like aging populations, advancements in healthcare technology, and a rising preference for home-based care by patients and their families (McKinsey 2022). Additionally, home healthcare is typically a more cost-effective option compared to hospital care, as it involves fewer resources (Singh et al. 2022).

Routing and scheduling caregivers for home visits is a complex but essential task for operational smoothness in home healthcare. Firstly, the process involves carefully balancing the needs of patients, caregivers, and the agency, encompassing timely care delivery, favorable working conditions, and cost-efficiency. Secondly, the complexity arises from the high number of decisions, which make it difficult to find optimal solutions within reasonable computation times. Thirdly, the uncertainty in patient conditions and traffic further complicates the problem.

In this work, we study the problem of simultaneously deciding on patient selection, assignment, routing and scheduling over a multi-day planning horizon with uncertain travel and service times. We find robust schedules that ensure timely care and prevent overtime, even with a worst-case realization of service and travel times. The conservatism of the schedules can be controlled through the utilization of uncertainty budgets, as proposed by Bertsimas and Sim (2004). Our approach takes into account a wide range of optimization criteria including patient time windows, number of required visits, caregiver availability, compatibility, continuity of care, and profit. The main contributions of this study can be summarized as follows:

Robustness. Building upon previous studies that primarily emphasized robustness concerning overtime, our research extends this focus to address the crucial element of ensuring timely patient care. The computational complexity increases as our approach requires meeting all patients' time windows in each worst-case scenario.

Health-Outcome Prioritizing Modeling Scheme. We introduce a systematic modeling scheme for routing and scheduling in home healthcare. Previous studies have implicitly prioritized multiple criteria in various forms, as can be seen in the literature reviews by Fikar and Hirsch (2017) and Cissé et al. (2017). In contrast, we strictly differentiate between health-related criteria (modeled as hard constraints) and efficiency-related criteria (optimized in the objective function). This clear distinction ensures that all accepted patients receive the best possible treatment regarding factors influenced by operational planning.

Optimal Solutions with Branch and Price. We propose a nested Dantzig-Wolfe decomposition to tackle the complex structure of home healthcare routing and scheduling (HHCR&S). Specifically, we decompose the problem into smaller subproblems for each caregiver, and subsequently into subproblems for each day. Based on this nested structure, we present the first branch-and-price method to solve a robust HHCR&S problem. Notably, this method also represents one of the first branch-and-price approaches for multi-period HHCR&S.

Case Study. Using real data from a home healthcare agency in Dallas, Texas, our computational study demonstrates that the nested branch-and-price procedure yields optimal robust solutions for multi-day problem instances of realistic size within reasonable runtimes. While non-robust solutions are found within seconds for most instances, robustness increases runtimes to several minutes. Moreover, compared to parameter settings lacking uncertainty or time windows, introducing a medium level of robustness and one-hour time windows leads to an average profit loss of 14.20%, primarily due to reduced patient loads per caregiver per day. Nonetheless, incorporating buffer time allows for low-cost robustification of some weekly schedules. Notably, our findings underscore that revenue takes precedence over geographic considerations in acceptance and assignment decisions.

The remainder of this chapter is organized as follows. Section 4.2 reviews related optimization problems in the HHCR&S literature. Section 4.3 presents a robust mixed-integer linear model for the problem of weekly HHCR&S. In Section 4.4, we develop a branch-and-price algorithm to solve the problem efficiently also for larger problem instances. In Section 4.5, we conduct a computational

study using real-world data from a home healthcare agency based in Dallas, Texas.

4.2. Literature Review

For a comprehensive survey of the HHCR&S literature, we refer to Cissé et al. (2017), Fikar and Hirsch (2017), and Di Mascolo et al. (2021). This review focuses on two streams that are most relevant to our research: In Section 4.2.1, we present an overview of exact solution methods for HHCR&S problems. In Section 4.2.2, we discuss robust optimization approaches in the presence of uncertain travel and/or service times. Our contribution to the literature lies in expanding the problem scope within each stream and innovatively combining both streams of research.

4.2.1. Exact Solutions

To find optimal solutions for larger instances of HHCR&S problems, exact solution methods leverage the decomposable problem structure. Specifically, for a given assignment of patients to caregivers, the routing and scheduling subproblems for each caregiver can be solved more efficiently. Based on this observation, two well-known exact methods have been applied in the literature: Logic-based Benders decomposition (LBBD) and Branch and Price (B&P).

LBBD. HHCR&S can be elegantly solved using Logic-based Benders Decomposition. This approach involves dividing the problem into a master problem that assigns patients to caregivers, and subproblems that determine the routing and scheduling for each caregiver.

The pioneering work by Heching and Hooker (2016) introduces LBBD to HHCR&S. Specifically, they study a multi-period problem aiming to maximize the number of patients served, while considering time windows and continuity of care. They address the master problem through mixed-integer linear programming and solve the subproblems by constraint programming. In subsequent work by Heching et al. (2019), they substantially enhance efficiency by introducing a time window subproblem relaxation, classical assignment relaxation, and multi-

commodity flow relaxation into the master problem; and introduce a variant of LBBD called branch and check. Grenouilleau et al. (2020) propose a new algorithm to handle the subproblems and present a novel LBBD formulation featuring variables representing pre-generated visit patterns. While they do not aim for exact solutions, their matheuristic focuses on efficiency. In contrast, Naderi et al. (2023) recently achieved an exact solution for a robust multi-period HHCR&S problem using LBBD. Their approach ensures robustness regarding overtime, but does not account for time windows. They minimize fixed caregiver costs and overtime costs. Notably, their LBBD approach differs from Heching et al. (2019) due to the inclusion of a nurse eligibility feature. Moreover, their subproblems are optimization problems rather than feasibility problems. Consequently, they additionally developed custom Benders optimality cuts.

Notably, the inherent objective function of the LBBD master problem does not account for subproblem variables. For example, the travel cost is dependent on the routes between patients and cannot be solely determined by assigning patients to caregivers. Even though the master problem's objective function can be modified to incorporate objective functions based on subproblem variables, if the original problem's objective function aims to minimize travel or wage costs, the LBBD master problem may still not serve as the most effective guide to achieving an optimal overall solution.

B&P. In the context of branch and price, the master problem selects either weekly or daily routes for each caregiver from a subset of all feasible solutions. The pricing problem, meanwhile, generates additional routes with better objective function values, if possible, and adds them to the pool of feasible solutions in the master problem. To the best of our knowledge, Rasmussen et al. (2012) are the first to employ a branch-and-price algorithm to a HHCR&S problem, maximizing the number of patients served for a single-period planning horizon. In contrast, Trautsamwieser and Hirsch (2014) address a multi-period context and minimize total working time. Yuan et al. (2014) and Yuan et al. (2015) tackle the minimization of total cost for a single-period planning horizon, accounting for stochastic service times and late arrivals, using branch and price. Liu et al. (2017) and Liu et al. (2019) refine this approach to minimize travel costs and penalties for unvisited patient, with the former considering lunch breaks and

the latter incorporating stochastic service and travel times. Qiu et al. (2022) adds synchronization constraints, using a strengthened branch-and-price-and-cut algorithm. Similarly, Tanoumand and Ünlüyurt (2021) concentrates on travel cost minimization with different vehicle types. Alkaabneh and Diabat (2023) explore a multi-objective scenario, aiming to minimize cost and maximize patient-caregiver compatibility, combining branch and price with a two-stage metaheuristic.

Recently, Yin et al. (2023) tailored an efficient branch-and-price-and-cut algorithm for a single-period planning problem, considering dispatching and transport costs, as well as incompatibility costs, and skill and time window constraints. Their B&P algorithm involves a bounded bidirectional labelling algorithm and heuristics to solve the subproblems, as well as subset-row inequalities. These studies collectively highlight the efficiency of branch-and-price approaches in addressing the complexities of HHCR&S.

While some of these approaches incorporate stochastic travel and service times, none considers robustness with respect to unfavorable realizations. In related problems, such as robust vehicle routing problems with time windows, prior research has demonstrated the effectiveness of branch and price (Pecin et al. 2017, Munari et al. 2019).

Our contributions to exact solution methods for HHCR&S are twofold. First, we build upon a problem framework similar to Heching et al. (2019), extend it to robustness under uncertain service and travel times, and incorporate the objective of minimizing both travel and wage costs while simultaneously maximizing the number of accepted patients. As we consider minimizing travel cost a substantial objective, we employ B&P as our solution approach. Second, our contributions to the B&P literature encompass two key aspects: introducing robustness for travel and service times, and extending solutions to a multi-day planning horizon.

4.2.2. Robust Home Healthcare Planning

Disruptions such as traffic delays and changing patient needs pose recurring challenges to timely service delivery. While the prevailing literature in home healthcare focuses on deterministic routing and scheduling, recent years have witnessed a growing research interest in stochastic and robust planning. Given the direct impact of both timely care and overtime on patient health outcomes (Aiken

et al. 2002, Bae and Fabry 2014, Hermida et al. 2016), our study aims to ensure punctuality not only on average but also in worst-case scenarios. In the following, we present a comprehensive review of robust approaches for HHCR&S. Our examination aligns with the cardinality-constrained robustness framework (or Γ-robustness) introduced by Bertsimas and Sim (2004). For fuzzy approaches, we point to Fathollahi-Fard et al. (2020a) and Shiri et al. (2023). Regarding robust methods with discrete scenario sets, which are computationally easier to integrate, we refer to Naji et al. (2017), Fathollahi-Fard et al. (2021), Fathollahi-Fard et al. (2022), and Shiri et al. (2023).

Initial robust optimization studies addressed uncertain visit demand (Carello and Lanzarone 2014, Carello et al. 2017), which clearly is important for the assignment, but does not ensure timely care. Pioneering cardinality-constrained uncertainty in travel and service times, Shi et al. (2019) tackle a single-period home healthcare scheduling problem, minimizing costs under time windows and skill requirements. They employed a recursive approach to linearize the adversarial problem. They find optimal solutions with a MILP solver and additional heuristic solutions with Simulated Annealing, Tabu Search and Variable Neighborhood Search. Makboul et al. (2021) consider a similar single-period problem, but additionally take lunch breaks into account. They solve the problem using a MILP solver. Shahnejat-Bushehri et al. (2021) extend the problem to multiple time windows, synchronization, and precedence constraints. They do not linearize the adversarial problem using the recursive approach, but enumerate all extreme points and solve the problem with metaheuristic algorithms. Fathollahi-Fard et al. (2020b) are the first to address a multi-period problem. They minimize total costs and carbon emissions, integrate robustness, and also solve the problem with metaheuristic algorithms. Finally, Naderi et al. (2023) solve the first multi-period robust staffing, assignment, routing and scheduling problem exactly. Their goal is to minimize fixed costs and overtime costs. They employ dualization to linearize the adversarial problem of maximizing overtime, and do not consider time windows. They use pre-generated visit patterns and a novel logic-based Benders branching-decomposition algorithm for exact solutions.

This work contributes to cardinality-constrained robust home healthcare routing and scheduling approaches as follows. Firstly, we ensure timely care within designated time windows on a multi-period planning horizon. Secondly, we con-

duct an analytical comparison of two linearization techniques for handling the robustness constraints. Thirdly, this study introduces a novel branch-and-price algorithm to find robust home healthcare routes and schedules, efficiently solving realistic problem instances to optimality.

4.3. Robust Routing and Scheduling of Home Healthcare Workers

4.3.1. Problem Statement

Given a set of patients and a set of caregivers, we address the problem of simultaneously determining the acceptance of new patients, assigning patients to caregivers, and creating daily routes and schedules for a set of patients and a set of home health caregivers, over a multi-day planning horizon. We propose a new classification scheme for this multi-criteria problem whose objective is to optimize cost and preference considerations. All criteria related to health outcome, such as continuity of care, punctuality, and compatible assignments, are met as hard constraints to ensure the best possible health outcomes. We note that quality measures related to health outcomes are also the primary determinant for home healthcare ratings (CMS 2020). Specifically, our model takes into account the following criteria.

4.3.1.1. Cost-Related Criteria

Maximum Profit. To determine the total profit, we calculate the revenue generated from treating patients and subtract wage costs and travel costs.

4.3.1.2. Health-Outcome Related Criteria

Patient Time Windows. A majority of patients in home healthcare require assistance with activities of daily living, including bathing, dressing, toileting, transferring, eating, and administering medications (Caffrey et al. 2011, Osakwe et al. 2019). Delays in these services can harm the patient's safety. Moreover, improper timing of certain medications, such as blood pressure medications or

anti-diabetic medications, can result in serious health complications and hospitalizations (Hermida et al. 2016). Therefore, we ensure that all patient time windows must be met even in a worst-case scenario of travel and service times.

Overtime. Overtime not only negatively impacts the physical and mental health of caregivers, but also the safety of patients and their health outcomes, see, e.g., Aiken et al. (2002), Bae and Fabry (2014). Therefore, we ensure that caregivers have no overtime, even in a worst-case scenario of travel and service times.

Continuity of Care. Interpersonal continuity of care not only builds trust and increases patient satisfaction, it also contributes to improved health outcome and lower morbidity (Cabana et al. 2004, Chan et al. 2021). This may be due, e.g., to a better understanding of the patient's medical history and treatment plan. We therefore enforce strict continuity of care, i.e., a patient is only accepted by the agency if he or she can be assigned to exactly one caregiver over the episode of care.

Compatible Assignments. Patient assignments must take into account the compatibility of the caregivers' skill levels, as well as other factors that may affect compatibility, such as gender and language.

4.3.2. Model Formulation

Indices. The planning horizon consists of multiple days, denoted by $d \in \mathcal{D}$. The available caregivers are denoted by $k \in \mathcal{K}$. There are two types of patient requests: new patient requests for treatment than can be admitted or rejected, denoted by $i \in \mathcal{I}^{\mathcal{R}}$, and patients who must receive treatment, denoted by $i \in \mathcal{I}^{\mathcal{E}}$, including existing patients with ongoing treatments from the previous planning period. Patients in the set $\mathcal{I}^{\mathcal{E}}$ are pre-assigned to the same caregiver and days as in the previous planning period to ensure strict continuity of care. For each caregiver k and day d, these patients compose the set $\mathcal{I}^{\mathcal{E}}_{k,d}$. All patients are grouped together in the set $\mathcal{I} := \mathcal{I}^{\mathcal{E}} \cup \mathcal{I}^{\mathcal{R}}$, and the total number of patients in $\mathcal{I}$ is n. We denote $N := \{1, \ldots, n\}$ and $N_0 := \{0\} \cup N$. The depot, which is

the starting and ending point for all caregivers, is indexed by $i \in \{0, n+1\}$, respectively.

Parameters. Caregiver k is available on day d during a working time window of $\left[\underline{W}_{kd}, \bar{W}_{kd}\right]$. Each patient i has a time window $[\underline{T}_{id}, \bar{T}_{id}]$ on day d. Moreover, patient i requires exactly v_i visits within the planning horizon, with at least d_i days between any two visits. A compatibility matrix $C \in \{0,1\}^{n \times |\mathcal{K}| \times |\mathcal{D}|}$ indicates which patients are compatible with which caregivers and on which days. Specifically, caregiver k can treat patient i on day d if and only if $C^i_{kd} = 1$. The revenue generated by caregiver k when treating patient i is denoted by R_{ki}. Furthermore, we take into account per time unit wage costs C^W_k and travel costs $C^{T^d}_{ij}$ between two locations i and j.

Uncertainty. We aim to find robust schedules in the presence of uncertain travel times $\tilde{t} \in \mathbb{R}_+^{(n+2)\times(n+2)}$ and uncertain service times $\tilde{p} \in \mathbb{R}^n_+$. For each patient $i \in \mathcal{I}$, we are given the expected service time $\bar{p}_i$ and the maximum positive deviation $\hat{p}_i$. Similarly, for each transfer from location i to location j, we are given the expected travel time $\bar{t}_{ij}$ and the maximum positive deviation $\hat{t}_{ij}$. The uncertainty budget for travel times is denoted by $\Gamma^t \in \mathbb{N}_0$, and the uncertainty budget for service times is denoted by $\Gamma^p \in \mathbb{N}_0$. Following the polyhedral uncertainty sets proposed in Bertsimas and Sim (2004), with the worst-case being achieved at the right-hand side of the uncertainty range, the uncertainty sets for travel and service times are defined as $\mathcal{U}^p := \{p : p_i = \bar{p}_i + \gamma^p_i \hat{p}_i, \sum_{i=1}^n \gamma^p_i \leq \Gamma^p, 0 \leq \gamma^p_i \leq 1 \, \forall i \in \{1, \ldots, n\}\}$ and
$\mathcal{U}^t := \left\{t : t_{ij} = \bar{t}_{ij} + \gamma^t_{ij}\hat{t}_{ij}, \sum_{i=0}^{n+1}\sum_{i=0}^{n+1} \gamma^t_{ij} \leq \Gamma^t, 0 \leq \gamma^t_{ij} \leq 1 \, \forall i, j \in [n+1]_0\}\right\}$, respectively. We refer to each element in the uncertainty sets of travel and service times as a travel or service time scenario, respectively. In particular, all uncertainty is assumed to be uncorrelated. The uncertainty budgets control the level of conservatism, and will be shown to correspond to the daily number of patients (transfers) per caregiver that require the maximum absolute deviation from their expected service (travel) time.

Variables. We use binary variables x^{kd}_{ij} to determine whether caregiver $k \in \mathcal{K}$ on day $d \in \mathcal{D}$ treats patient $j \in \mathcal{I} \cup \{0, n+1\}$ immediately after patient $i \in$

$\mathcal{I} \cup \{0, n+1\}$. We use binary variables $\delta_{ki} \in \{0,1\}$ to determine whether patient $i \in \mathcal{I}$ is assigned to caregiver $k \in \mathcal{K}$. Moreover, with $M := \underline{T}_{id} - \bar{T}_{jd}$, we find the scenario-dependent start time of patient $i \in \mathcal{I}$ on day d as:

$$s_{i,d}(x,p,t) := \min \left\{ \begin{array}{l} \underline{T}_{id}, \\ \max \left\{ \begin{array}{l} \sum\limits_{k \in \mathcal{K}} (\underline{W}_{kd} + t_{0i})\, x_{0i}^{kd}, \\ \max\limits_{j \in \mathcal{I}} \left\{ s_j + (p_j + t_{ji}) \sum\limits_{k \in \mathcal{K}} x_{ji}^{kd} + M(1 - \sum\limits_{k \in \mathcal{K}} x_{ji}^{kd}) \right\} \end{array} \right\} \end{array} \right\}$$

The first term of the minimization function indicates that patient i is available for treatment starting with the lower bound $\underline{T}_{id}$ of his or her time window. The second term in the minimization function states that caregiver k starts working at time $\underline{W}_{kd}$, and is available to treat patient i only after completing treatment of the previous patient and traveling to patient i.

Similarly, with $M := \underline{W}_{kd} - \bar{T}_{id}$, we define the scenario-dependent return time of caregiver k to the depot on day d as follows:

$$s_{kd}^{n+1}(x,p,t) := \min \left\{ \begin{array}{l} \underline{W}_{kd}, \\ \max\limits_{i \in \mathcal{I}} \left\{ s_{i,d}(x,p,t) + (p_i + t_{i,n+1}) x_{i,n+1}^{kd} + M(1 - x_{i,n+1}^{kd}) \right\} \end{array} \right\}$$

The first term of the minimization function ensures that caregiver k cannot return to the depot before starting his or her shift. The second term in the minimization function states that the caregiver returns to the depot only after completing the treatment of the last patient and traveling to the depot.

We now consider the following Robust Home Healthcare Routing and Scheduling Problem (R-HHCR&S).

$$\max \sum_{k\in\mathcal{K}} \left(\sum_{i\in\mathcal{I}} R_{ki} v_i \delta_{ki} - \sum_{d\in\mathcal{D}} \sum_{i,j\in\mathcal{I}} C^T_{dij} x^{kd}_{ij} - C^W_k \sum_{d\in\mathcal{D}} \left(s^{n+1}_{kd} - \underline{W}_{kd} \right) \right) \tag{4.1}$$

$$s.t. \max_{\substack{p\in\mathcal{U}^p \\ t\in\mathcal{U}^t}} s_{id}(x,p,t) \leq \bar{T}_{id} \qquad \forall d\in\mathcal{D}, i\in\mathcal{I} \tag{4.2}$$

$$\max_{\substack{p\in\mathcal{U}^p \\ t\in\mathcal{U}^t}} s^{n+1}_{kd}(x,p,t) =: s^{n+1}_{kd} \leq \bar{W}_{kd} \qquad \forall k\in\mathcal{K}, d\in\mathcal{D} \tag{4.3}$$

$$\sum_{d\in\mathcal{D}} \sum_{j\in\mathcal{I}} x^{kd}_{ij} = v_i \delta_{ki} \qquad \forall k\in\mathcal{K}, i\in\mathcal{I} \tag{4.4}$$

$$\sum_{k\in\mathcal{K}} \delta_{ki} \leq 1 \qquad \forall i\in\mathcal{I} \tag{4.5}$$

$$\sum_{i\in N_0} x^{kd}_{ji} = 1 \qquad \forall k\in\mathcal{K}, i\in\mathcal{I}^{E|_{k,d}} \tag{4.6}$$

$$\sum_{j\in\mathcal{I}} x^{kd}_{ij} \leq C^{kd}_i \qquad \forall k\in\mathcal{K}, d\in\mathcal{D}, i\in\mathcal{I}^{\mathcal{R}} \tag{4.7}$$

$$\sum_{k\in\mathcal{K}} \sum_{j\in\mathcal{I}} x^{kd}_{ij} + \sum_{k\in\mathcal{K}} \sum_{\mu\in[d_i]_0} \sum_{j\in\mathcal{I}} x^{kd+\mu}_{ij} \leq 1 \qquad \forall d\in\mathcal{D}, i\in\mathcal{I}^{\mathcal{R}} \tag{4.8}$$

$$\sum_{j\in\mathcal{J}} x^{kd}_{ji} - \sum_{j\in\mathcal{J}} x^{kd}_{ij} = 0, \quad \sum_{i\in\mathcal{J}} x^{kd}_{ii} = 0 \qquad \forall k\in\mathcal{K}, d\in\mathcal{D}, i\in\mathcal{I} \tag{4.9}$$

$$\sum_{i\in[n+1]} x^{kd}_{0,i} = 1, \quad \sum_{i\in N_0} x^{kd}_{i,n+1} = 1 \qquad \forall k\in\mathcal{K}, d\in\mathcal{D} \tag{4.10}$$

$$x^{kd}_{ij}, \delta_{ki} \in \{0,1\} \qquad \forall k\in\mathcal{K}, d\in\mathcal{D}, i,j\in[n+1]_0 \tag{4.11}$$

Objective Function (4.1) maximizes the total profit, calculated by subtracting travel and wage costs from revenue. Constraints (4.2) ensure robust schedules by guaranteeing that patient time windows are met even in a worst-case scenario of travel and service times. Similarly, Constraints (4.3) create robust schedules by ensuring that the caregivers' work hours are not exceeded even in a worst-case scenario of travel and service times. We call the maximization term within Constraints (4.2) and (4.3) the *adversarial problems*. Due to their nonlinear nature, (4.1) – (4.11) is not amenable to MILP solvers as it is currently formulated.

Constraints (4.4) ensure that each accepted patient is visited the exact number

of required times. By assigning each patient to at most one caregiver, Constraints (4.5) enforce strict continuity of care. Constraints (4.6) ensure that all patients in the set $\mathcal{I}^{\mathcal{E}}$ are visited by their pre-assigned caregiver on the designated day. Notably, adjusting the time windows allows for the straightforward fixing of specific visits to particular times. Constraints (4.7) specify individual patient-caregiver compatibility. Constraints (4.8) enforce a minimum number of days between any two treatments for a patient.

Constraints (4.9) are the well-known flow-conservation constraints. Moreover, Constraints (4.10) ensure that each caregiver starts and ends at a central depot. Together, the last two constraints ensure feasible routes, with Constraints (4.7) eliminating subtours.

The model contains $\mathcal{O}\left(\mathcal{K}| \cdot |\mathcal{D}| \cdot n^2\right)$ binary variables, $\mathcal{O}\left(|\mathcal{K}| \cdot |\mathcal{D}| \cdot n\right)$ continuous variables and $\mathcal{O}(n \cdot |\mathcal{K}| \cdot |\mathcal{D}|)$ constraints. After linearizing Constraints (4.2) and (4.3) it becomes a MILP.

4.3.3. Adversarial Problems

The *dualization scheme* proposed by Bertsimas and Sim (2004) has classically been used for linearizing adversarial problems in cardinality-constrained robust optimization. However, due to the nested definition of service start times, its application to our adversarial problems is not straightforward. To address this issue, various alternative approaches have been proposed in the literature, such as by Agra et al. (2013) and by Lee et al. (2012). One of the most efficient approaches currently available is the *recursive scheme* introduced by Munari et al. (2019) for a robust vehicle routing problem with time windows.

In the following, we compare the effectiveness of the classical dualization scheme, based on a novel reformulation of service start times, with the effectiveness of the recursive scheme for linearizing Constraints (4.2) and (4.3). While the novel reformulation provides a deeper understanding of robust start times and how to choose the uncertainty budgets, our theoretical analysis demonstrates that the recursive scheme remains the more efficient method for linearizing the adversarial problems in R-HHCR&S.

4.3.3.1. Dualization Scheme

To apply the classical dualization scheme as proposed by Bertsimas and Sim (2004), we reformulate the service start times. In particular, we replace the binary variable x with a new binary variable $z_{j'j}^{k,d,i} \in \{0,1\}$, which additionally indicates the sequence of patients treated by caregiver k on day d. Specifically, $z_{j'j}^{kdi}$ takes a value of 1 if and only if the ith patient treated by caregiver k on day d is patient j, and the patient treated before j was patient j'; 0 otherwise. We now shift the focus from the start time s_{id} of patient i to the start time s_{kd}^i of the ith patient treated by caregiver k on day d. This start time can be calculated as follows (see Appendix A.3.1):

$$\min s_{kd}^i \tag{4.12}$$

$$\text{s.t. } s_{kd}^i \geq \underline{W}_{kd} + \max_{\substack{p\in\mathcal{U}^p \\ t\in\mathcal{U}^t}} \left(\sum_{i'=1}^{i-1} \sum_{j'=0}^{n} \sum_{j=1}^{n} (t_{j'j} + p_j) z_{j'j}^{k,d,i'} + \sum_{j'=0}^{n} \sum_{j=1}^{n} t_{j'j} z_{j'j}^{k,d,i} \right) \tag{4.13}$$

$$s_{kd}^i \geq \sum_{j'=0}^{n} \sum_{j=1}^{n} \underline{T}_{jd} z_{j'j}^{k,d,i^*} + \max_{\substack{p\in\mathcal{U}^p \\ t\in\mathcal{U}^t}} \sum_{i'=i^*+1}^{i} \sum_{j'=0}^{n} \sum_{j=1}^{n} (p_{j'} + t_{j'j}) z_{j'j}^{k,d,i'} \qquad \forall i^* \in [i] \tag{4.14}$$

Model (4.12) – (4.14) ensures that the start time of each patient equals the maximum of the earliest availability of all previous visits (including the depot) plus a worst-case scenario of service and travel times to the current patient. This formulations allows us to now apply the dualization scheme (see Appendix A.3.1 for the derivation).

Let n_{max} denote the maximum number of patients treated by any caregiver on any day. The linearization of the robustness Constraints (4.2) and (4.3) using the proposed dualization approach requires in total $n^2 \cdot n_{max} \cdot |\mathcal{K}| \cdot |\mathcal{D}|$ binary variables, $\mathcal{O}\left(n^2 \cdot (n_{max})^2 \cdot |\mathcal{K}| \cdot |\mathcal{D}|\right)$ continuous variables and $\mathcal{O}\left(n^2 \cdot |\mathcal{K}| \cdot |\mathcal{D}| \cdot (n_{max})^2\right)$ new constraints. We note that if the uncertainty budget for travel time $\Gamma^t \geq 3$, then the proposed dualization scheme is more efficient compared to the linearization approach given by Agra et al. (2013).

4.3.3.2. Recursive Scheme

Next, we apply the recursive scheme proposed by Munari et al. (2019) to linearize the problem. Let $s^d_{(i,\gamma^p,\gamma^t)}$ denote the service start time of patient i on day d, where $\gamma^p \in \{0,\dots,\Gamma^p\}$ and $\gamma^t \in \{0,\dots,\Gamma^t\}$ denote the units of maximum positive deviation that have already been consumed from the uncertainty budgets. Then we can replace the robustness Constraints (4.2) and (4.3) by

$$s^d_{i,\gamma^p,\gamma^t} \in [\underline{T}_{id}, \bar{T}_{id}] \quad \forall d \in \mathcal{D}, i \in \mathcal{I}, \gamma^p \in [\Gamma^p]_0, \gamma^t \in [\Gamma^t]_0 \tag{4.15}$$

$$s^d_{i,0,\zeta^t} \geq \sum_{k\in\mathcal{K}} \left(\underline{W}_{kd} + \bar{t}_{0i} + \zeta^t \hat{t}_{0i}\right) x^{kd}_{0i} \quad \forall d \in \mathcal{D}, i \in \mathcal{I} \, \forall \zeta^t \in \{0,1\} \tag{4.16}$$

$$\begin{aligned} s^d_{i,\gamma^p+\zeta^p,\gamma^t+\zeta^t} \geq\ & s^d_{j,\gamma^p,\gamma^t} + (\bar{p}_j + \zeta^p \hat{p}_j + \bar{t}_{ji} + \zeta^t \hat{t}_{ji}) \sum_{k\in\mathcal{K}} x^{kd}_{ji} \\ & + (\underline{T}_{id} - \bar{T}_{jd})(1 - \sum_{k\in\mathcal{K}} x^{kd}_{ji}) \\ & \forall i,j \in \mathcal{I}, \gamma^p \in [\Gamma^p]_0, d \in \mathcal{D}, \gamma^t \in [\Gamma^t]_0, \zeta^p, \zeta^t \in \{0,1\} : \\ & \gamma^p + \zeta^p \leq \Gamma^p, \gamma^t + \zeta^t \leq \Gamma^t \end{aligned} \tag{4.17}$$

and similarly, for avoiding overtime,

$$s^{n+1}_{kd} \in [\underline{W}_{kd}, \bar{W}_{kd}] \quad \forall k \in \mathcal{K}, d \in \mathcal{D} \tag{4.18}$$

$$\begin{aligned} s^{n+1}_{kd} \geq\ & s^d_{i,\Gamma^p-\zeta^p,\Gamma^t-\zeta^t} + (\bar{p}_i + \zeta^p \hat{p}_i + \bar{t}_{i,n+1} + \zeta^t \hat{t}_{i,n+1}) x^{kd}_{i,n+1} \\ & + (\underline{W}_{kd} - \bar{T}_{id})(1 - x^{kd}_{i,n+1}) \\ & \forall k \in \mathcal{K}, d \in \mathcal{D}, i \in \mathcal{I}, \zeta^p, \zeta^t \in \{0,1\} \end{aligned} \tag{4.19}$$

This leads to $\mathcal{O}\left(n \cdot |\mathcal{D}| \cdot |\Gamma^p| \cdot |\Gamma^t| + |\mathcal{K}| \cdot |\mathcal{D}|\right)$ new continuous variables, $\mathcal{O}\left(n^2 \cdot |\mathcal{D}| \cdot |\Gamma^p| \cdot |\Gamma^t|\right)$ new constraints, but no new binary variables. In contrast to the dualization scheme, the number of variables in the recursive scheme depends on the uncertainty budget.

4.3.3.3. Comparison

Based on the number of variables and constraints used in each approach, we finally compare the effectiveness of the dualization scheme and the recursive scheme to linearize our robustness constraints.

Firstly, the recursive scheme stands out as it requires significantly fewer binary variables, an order of n_{max} times less than the dualization scheme. Secondly, if $|\mathcal{K}| \cdot (n_{max})^2 > |\Gamma^p| \cdot |\Gamma^t|$, then the recursive scheme also requires fewer continuous variables and constraints.

Despite this, examining the reformulation of start times in (4.12) – (4.14) reveals that the uncertainty budgets limit the amount of uncertainty faced by each caregiver per day, and should be set accordingly. Therefore, w.l.o.g., we set the uncertainty budgets such that $\Gamma^p \leq n_{max}$ and $\Gamma^t \leq n_{max} + 2$, a choice that further enhances the efficiency of the recursive scheme. It is worth noting that if the budgets exceed the maximum number of patients visited and paths traveled by a caregiver on a given day, the problem simplifies to a deterministic problem in which it is assumed that all patient visits and traveled paths require the maximum positive deviation in addition to the expected time.

In conclusion, the recursive scheme proves more efficient in linearizing the robustness constraints in R-HHCR&S. As a result, we replace (4.2) – (4.3) with the linear inequalities (4.15) – (4.19).

4.4. Solution Method

To solve R-HHCR&S for larger instances, we exploit its decomposable problem structure. Specifically, the problem can be decomposed into subproblems for each caregiver, and each caregiver's subproblem can be further decomposed into daily routing and scheduling subproblems, as depicted in Figure 4.1. Our nested branch-and-price procedure, outlined in Figure 4.2, leverages this decomposition to recursively compute solutions across three levels.

At the first level, the restricted master problem combines given plans for all caregivers into a comprehensive weekly plan for the agency. Integral solutions for the selection of caregiver plans are ensured via a branch-and-bound procedure. In the second level, new weekly plans for each caregiver are generated, terminat-

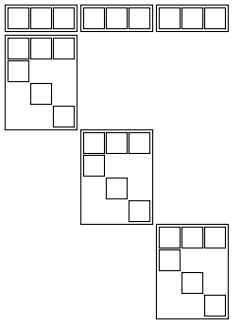

Figure 4.1. Two-fold decomposition of the problem structure. The small boxes represent daily routing and scheduling problems for each caregiver. The daily solutions for each caregiver are combined to create multi-day solutions, represented by the large boxes. The corresponding combination process ensures that the required number of treatments is met, and the minimum number of days between two visits is satisfied. The multi-day plans for each caregiver are further combined to ensure strict continuity of care and minimize overall cost.

ing either with a caregiver's weekly plan that potentially improves the agency's overall plan or with the proof that none exists. Weekly caregiver schedules are generated using the second-level relaxed master problem, which combines daily plans into a weekly schedule. Integral solutions for daily route selection are again ensured through a branch-and-bound procedure. This is accomplished using a branch-and-bound procedure with robust traveling salesman subproblems incorporating time windows, the acceptance decision being guided by a multi-Knapsack problem.

Overall, our nested branch-and-price procedure efficiently explores promising daily routes and weekly plans, combining them into a comprehensive agency plan. Notably, while nested decomposition approaches have been employed in related problems (Vanderbeck 2001, Hennig et al. 2012, Dohn and Mason 2013, Tilk et al. 2019, Zheng et al. 2021), to the best of our knowledge we are the first to apply this method to an HHCR&S problem.

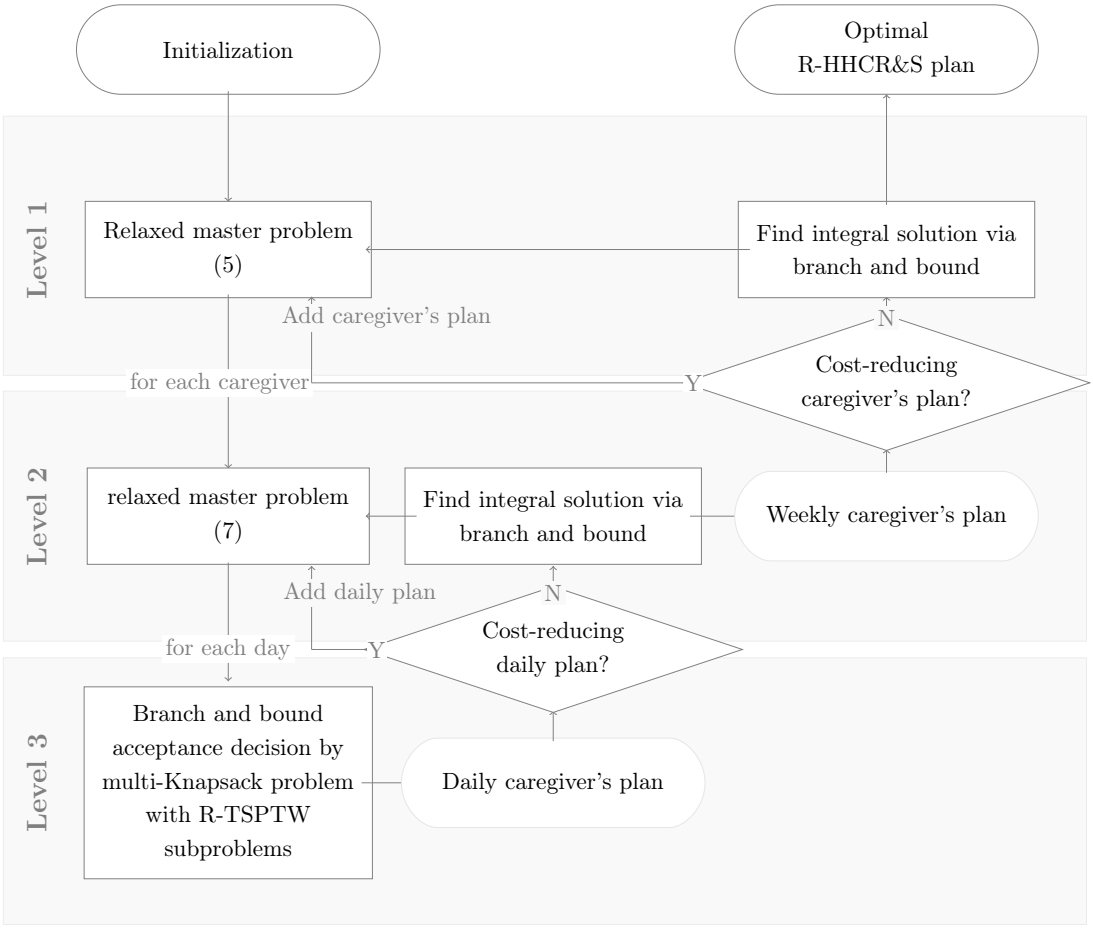

Figure 4.2. Flowchart of the nested branch-and-price procedure

4.4.1. Nested Dantzig-Wolfe Decomposition

4.4.1.1. First-Level Decomposition

We first decompose R-HHCR&S into a master problem and $|\mathcal{K}|$ pricing problems. Each pricing problem, associated with a fixed caregiver $k \in \mathcal{K}$, determines a weekly plan P that encodes a schedule and route for each day. We note that caregivers are heterogeneous in compatibility to patients, work times, wage cost, and revenue. The master problem then combines the weekly plans generated by the pricing problems in order to maximize profit, while ensuring that care is provided to all accepted patients.

Notation. We use a binary variable $\lambda_{k,P}$ to decide whether caregiver k is assigned to the weekly plan P. The set of all feasible weekly plans for caregiver k is denoted as $\mathcal{P}_k$. We use the binary parameter $a^i_{k,P}$ to indicate whether patient i is being treated by caregiver k in plan P. The revenue generated by a plan P executed by caregiver k is denoted by $Rev_{k,P}$, and similarly the travel cost $C^T_{k,P}$ and wage cost $C^W_{k,P}$.

Master Problem. The master problem is a variant of the set partitioning problem, in which elements are either accepted or rejected, and the accepted elements are partitioned into $|\mathcal{K}|$ subsets. Similar to other works such as Bard et al. (2014), it can be stated as follows:

$$\max \sum_{k\in\mathcal{K}} \sum_{P\in\mathcal{P}_k} \left(Rev_{k,P} - C^T_{k,P} - C^W_{k,P}\right) \lambda_{k,P} \tag{4.20}$$

$$\text{s.t.} \sum_{P\in\mathcal{P}_k} \lambda_{k,P} = 1 \qquad \forall k \in \mathcal{K} \tag{4.21}$$

$$\sum_{k\in\mathcal{K}} \sum_{P\in\mathcal{P}_k} a^i_{k,P}\lambda_{k,P} \leq 1 \qquad \forall i \in \mathcal{I}^{\mathcal{R}} \tag{4.22}$$

$$\lambda_{k,P} \in \{0,1\} \qquad \forall k \in \mathcal{K}, P \in \mathcal{P}_k \tag{4.23}$$

Objective Function (4.20) maximizes the total revenue. Constraints (4.22) ensure that each caregiver is assigned to exactly one plan. Constraints (4.23) ensure that new patients are either rejected or accepted and assigned to exactly one caregiver's plan.

Pricing Problems. The LP dual (A.39) – (A.41) of the master problem's LP relaxation is given in Appendix A.3.2.1. Let $v^* \in \mathbb{R}^{|\mathcal{K}|}$ and $w^* \in \mathbb{R}^{|\mathcal{I}^{\mathcal{R}}|}_+$ be a solution of the LP dual. We use the binary parameter a^i_R to indicate whether patient i is visited by caregiver k in route R. Then for each caregiver $k \in \mathcal{K}$, we must solve the pricing problem:

$$\max_{P \in \mathcal{P}_k} Rev_{k,P} - C^T_{k,P} - C^W_{k,P} - \sum_{i \in \mathcal{I}^{\mathcal{R}}} w^*_i \, a^i_{k,P} \tag{4.24}$$

The full pricing problem (A.47) – (A.60) is stated in Appendix A.3.2.2 and is seen to be a MILP with $\mathcal{O}\left(|\mathcal{I}^{\mathcal{R}} \cup \mathcal{I}^{\mathcal{E}}_k|^2 \cdot |\mathcal{D}|\right)$ binary variables and $\mathcal{O}\left(|\mathcal{I}^{\mathcal{R}} \cup \mathcal{I}^{\mathcal{E}}_k| \cdot |\mathcal{D}| \cdot |\Gamma^p| \cdot |\Gamma^t|\right)$ continuous variables. However, each pricing problem is further decomposable into days.

4.4.1.2. Second-Level Decomposition

We further decompose each caregiver k's pricing problem into a master problem and $|\mathcal{D}|$ pricing subproblems. The objective of each subproblem is to find a better route for day $d \in \mathcal{D}$. The master problem combines the daily routes into a weekly plan.

Notation. For any given caregiver k, we use the binary variable $\lambda_{d,r}$ to decide whether on day d he or she takes route r. The set of feasible routes on day $d \in \mathcal{D}$ is denoted as $\mathcal{R}_d$. The binary parameter $a^i_{d,r}$ indicates whether patient i is treated on day d in route r. The revenue generated by a route r executed by caregiver k is denoted by $Rev_{k,r}$, and similarly the travel cost on day d by $C^T_{d,r}$ and wage cost by $C^W_{k,r}$.

Master Problem.

$$\max \sum_{d\in\mathcal{D}}\sum_{r\in\mathcal{R}_d}\left(Rev_{d,r} - C^T_{d,r} - C^W_{k,r} - \sum_{i\in\mathcal{I}^{\mathcal{R}}}\frac{w^*_i}{v_i}a^i_{d,r}\right)\lambda_{d,r} \tag{4.25}$$

$$\text{s.t.} \sum_{r\in\mathcal{R}_d}\lambda_{d,r} = 1 \qquad \forall d\in\mathcal{D} \tag{4.26}$$

$$\sum_{d\in\mathcal{D}}\sum_{r\in\mathcal{R}_d} a^i_{d,r}\lambda_{d,r} \le v_i \qquad \forall i\in\mathcal{I}^{\mathcal{R}} \tag{4.27}$$

$$\sum_{r\in\mathcal{R}_d} v_i a^i_{d,r}\lambda_{d,r} - \sum_{\mu\in\mathcal{D}}\sum_{r\in\mathcal{R}_\mu} a^i_{\mu,r}\lambda_{\mu,r} \le 0 \qquad \forall d\in\mathcal{D}, i\in\mathcal{I}^{\mathcal{R}} \tag{4.28}$$

$$\sum_{r\in\mathcal{R}_d} a^i_R\lambda_{d,r} + \sum_{\mu\in[d_i]_0}\sum_{r\in\mathcal{R}_d} a^i_R\lambda^{d+\mu,r} \le 1 \qquad \forall d\in\mathcal{D}, i\in\mathcal{I}^{\mathcal{R}} \tag{4.29}$$

$$\lambda_{d,r}\in\{0,1\} \qquad \forall d\in\mathcal{D}, r\in\mathcal{R}_d \tag{4.30}$$

Objective Function (4.25) maximizes the total profit and penalizes the acceptance of new patients. Constraints (4.27) ensure that the caregiver takes exactly one route on each day. Constraints (4.28) and (4.29) ensure that new patients are either rejected or receive the required number of treatments. Constraints (4.30) ensure a minimum number of days between any two visits of a patient.

Pricing Subproblems. The LP dual (A.42) – (A.46) of the master problem's LP relaxation is given in Appendix A.3.2.1. Let $u^*\in\mathbb{R}^{|\mathcal{D}|}$, $z^*\in\mathbb{R}^{|\mathcal{I}^{\mathcal{R}}|}_{\ge 0}$, $y^*\in\mathbb{R}^{|\mathcal{D}|\times|\mathcal{I}^{\mathcal{R}}|}_{\ge 0}$, and $q^*\in\mathbb{R}^{|\mathcal{D}|\times|\mathcal{I}^{\mathcal{R}}|}_{\ge 0}$ be a solution of the LP dual. For better readability, we denote $\mathcal{I}_k := \mathcal{I}^{\mathcal{R}}\cup\mathcal{I}^{\mathcal{E}}_k$. Then we must solve the following pricing problem for caregiver k on day d:

$$\begin{aligned}\max_{r\in\mathcal{R}_d} & Rev_r - C^T_r - C^W_r \\ & - \left(\sum_{i\in\mathcal{I}^{\mathcal{R}}}\frac{w^*_i}{v_i} + \sum_{i\in\mathcal{I}^{\mathcal{R}}} z^*_i + \sum_{i\in\mathcal{I}^{\mathcal{R}}}\left(\sum_{\tilde{d}\in\mathcal{D}} y^*_{i\tilde{d}} - v_i y^*_{id}\right) + \sum_{i\in\mathcal{I}_k}\sum_{\substack{\mu\in[|\mathcal{D}|-d_i]:\\ \mu\in[d_i-d]_0}} q^*_{i,\mu}\right) a^i_r\end{aligned} \tag{4.31}$$

The corresponding MILP that represents the second-level pricing subproblem is

stated in Appendix A.3.2.2. It has $\mathcal{O}\left(|\mathcal{I}_k|^2\right)$ binary variables, $\mathcal{O}\left(|\mathcal{I}_k| \cdot |\Gamma^p| \cdot |\Gamma^t|\right)$ continuous variables, and $\mathcal{O}\left(|\mathcal{I}_k|^2 \cdot |\Gamma^p| \cdot |\Gamma^t|\right)$ constraints.

4.4.2. Column Generation

Generating all feasible daily routes $\mathcal{R}_{k,d}$ for (4.25) – (4.30) and all weekly plans $\mathcal{P}_k$ for (4.20) – (4.23) explicitly is impractical due to the typically high number of potential routes and plans. Therefore, we employ column generation to solve the linear relaxation of the master problems (Lübbecke and Desrosiers 2005, Desrosiers and Lübbecke 2011).

4.4.2.1. Initialization

Column generation begins with computing a subset $\tilde{\mathcal{P}}_k \subset \mathcal{P}_k$ of weekly plans for each caregiver k and a subset $\tilde{\mathcal{R}}_{k,d} \subset \mathcal{R}_{k,d}$ of routes for each day d, respectively. We initialize the subsets by devising two feasible solutions: The first solution entails visiting all existing patients on the pre-determined days while rejecting all new patients. The second solution employs a greedy heuristic to accept new patients.

The greedy heuristic systematically selects promising assignments as follows. An assignment consists of a patient $i \in \mathcal{I}^{\mathcal{R}}$, a compatible caregiver k, and a day pattern $D \subset \mathcal{D}$, with $|D| = v_i$ and the required number d_i of days between any two days in D. The value of this assignment is the revenue minus an estimation of the accompanying wage and travel cost: $R_{ki} - C_k^W \cdot \left(\min_{j \in \mathcal{I}_k \cup \{0\}} \bar{t}_{ij} + \bar{p}_i\right) - \min_{j \in \mathcal{I}_k \cup \{0\}} C^T_{dji}$. The wage cost for a new patient is approximated by the caregiver's wage cost multiplied by the expected service time, plus the minimum possible expected travel time to reach the patient. The travel cost is estimated as the minimum travel cost from any potential location to the new patient. Given an approximate value for each assignment, we iteratively choose the most promising assignment. We assign the new patient to the corresponding caregiver and day pattern, ensuring the corresponding daily routes remain feasible.

These subsets, $\tilde{\mathcal{P}}_k \subset \mathcal{P}_k$ and $\tilde{\mathcal{R}}_{k,d} \subset \mathcal{R}_{k,d}$, initialized by the reject-all-new-patients solution and the greedy-assignment solution, serve as initial columns for the respective restricted master problems. That is, the first-layer restricted master problem is derived by substituting $\mathcal{P}_k$ with $\tilde{\mathcal{P}}_k$ in the linear relaxation of

the master problem (4.25) – (4.30). Similarly, the second-layer restricted master problem is derived by replacing $\mathcal{R}_{k,d}$ with $\tilde{\mathcal{R}}_{k,d}$ in the linear relaxation of the master problem (4.20) – (4.23).

4.4.2.2. Pricing Problems

The restricted master problems only include a subset of feasible columns. Consequently, they yield solutions that are feasible for the linear relaxation of the original master problems but not necessarily optimal. We verify optimality through the pricing subproblems.

At the first level, we determine if the optimal solution of the restricted master problem remains optimal for the linear relaxation of the original master problem by determining whether there exists a weekly plan with a value exceeding the dual objective function value y_k^* for each caregiver k, as determined in (4.24). This assessment is carried out using the second-level branch-and-price procedure.

At the second level, we evaluate the optimality of the restricted master problem by checking for each day d whether there exists a route with a value greater than the dual objective function value u_d^* according to (4.31). It is worth noting that, owing to revenue maximization, our second-level pricing problems deviate from the common resource-constrained elementary shortest path pricing problems, and align with the more difficult robust longest path problem. To tackle the second-level pricing problems, we employ a simple branch-and-bound procedure, systematically accepting new patients, as outlined in the remainder of this section.

Branch and Bound. The second-level branch-and-bound procedure is strongly guided by upper bounds $\bar{v}_{kd}(i)$ on the value generated by caregiver k visiting patient $i \in \mathcal{I}_k$ on day d, calculated as:

$$\begin{aligned}\bar{v}_{kd}(i) := R_{ki} - C_k^W \cdot \left(\min_{j \in \mathcal{I}_k \cup \{0\}} \bar{t}_{ij} + \bar{p}_i \right) - \min_{j \in \mathcal{I}_k \cup \{0\}} C_{dji}^T \\ - \mathbb{1}_{i \in \mathcal{I}^{\mathcal{R}}} \left(\frac{w_i^*}{v_i} + z_i^* + \left(\sum_{\mu \in \mathcal{D}} y_{i\mu}^* - v_i y_{id}^* \right) + \sum_{\mu \in [|\mathcal{D}| - d_i] : \mu \in [d_i - d]_0} q_{i\mu}^* \right)\end{aligned} \tag{4.32}$$

The value $\bar{v}_{kd}(i)$ is derived from the revenue obtained from visiting patient i minus a lower bound on wage and travel costs incurred by by the visit to patient i, alongside the dual values delineated in (4.32). It is noteworthy that the dual values strongly inform the upper bound.

First, the upper bounds are used for a pre-selection of patients, where new patients with a value $\bar{v}_{kd}(i) \leq 0$ are excluded from consideration in the branch-and-bound procedure. Not visiting them leads to a higher value in the second-level pricing problem. Secondly, the upper bounds $\bar{v}_{kd}(i)$ allow for the computation of an upper bound on the objective function value of the second-level pricing problem. Let n_k be an upper bound on the number of patients that caregiver k can visit in a single day, and binary variables δ_i indicate whether caregiver k visits patient $i \in \mathcal{I}^{\mathcal{R}}$ on day d or not. Then the second-level pricing problem (4.31) is upper bounded by the solution of the following multi-knapsack problem:

$$\max \sum_{i \in \mathcal{I}^{E|_{k,d}}} \bar{v}_{kd}(i) + \sum_{i \in \mathcal{I}^{\mathcal{R}}} \bar{v}_{kd}(i) \cdot \delta_i \tag{4.33}$$

$$\text{s.t.} \sum_{i \in \mathcal{I}^{\mathcal{R}}} \left(\min_{j \in \mathcal{I}_k \cup \{0\}} \bar{t}_{ij} + \bar{p}_i \right) \delta_i \leq \bar{W}_{kd} - \underline{W}_{kd} - \sum_{i \in \mathcal{I}^{E|_{k,d}}} \left(\min_{j \in \mathcal{I}_k \cup \{0\}} \bar{t}_{ij} + \bar{p}_i \right) \tag{4.34}$$

$$\sum_{i \in \mathcal{I}^{\mathcal{R}}} \delta_i \leq n_{k,d} - |\mathcal{I}^{E|_{k,d}}| \tag{4.35}$$

$$\delta_i \in \{0, 1\} \; \forall i \in \mathcal{I}^{\mathcal{R}} \tag{4.36}$$

In this problem, Objective Function (4.33) aims to maximize the sum of upper bounds $\bar{v}_{kd}(i)$ for the value generated by visiting new patients. Constraint (4.34) ensures that a lower bound on the total workload does not exceed the caregiver's working hours, while Constraint (4.35) ensures that at most n_k patients are visited in total.

Now, if the upper bound provided by (4.33) – (4.36) does not surpass the dual value u_d^* or if the problem is infeasible, it provides evidence that no improving column exists. However, if it exceeds the dual value, the assignment determined by the multi-knapsack problem may or may not lead to an improving column. To verify, we solve the corresponding robust traveling salesman subproblems with time windows and wage cost (R-TSPTW) for caregiver k to determine whether

the assignment $\mathcal{I}_{TSP} := \mathcal{I}^{E|k,d} \cup \{i \in \mathcal{I}_k : \delta_i^* = 1\}$ for the current knapsack solution δ^* indeed yields a better column:

$$\min \sum_{i,j \in \mathcal{I}_{TSP}} C_{dij}^T x_{ij} + C_k^W \left(s_{n+1} - \underline{W}_{kd}\right) \tag{4.37}$$

$$\text{s.t. } s_{i,\gamma^p,\gamma^t} \in [\underline{T}_{id}, \bar{T}_{id}] \qquad \forall i \in \mathcal{I}_{TSP}\,, \gamma^p \in [\Gamma^p]_0\,, \gamma^t \in [\Gamma^t]_0 \tag{4.38}$$

$$s_{i,0,\zeta^t} \geq \left(\underline{W}_{kd} + \bar{t}_{0i} + \zeta^t \hat{t}_{0i}\right) x_{0i} + (\underline{W}_{kd} - \bar{T}_{id})(1 - x_{i,n+1}) \qquad \forall i \in \mathcal{I}_{TSP}\, \zeta^p, \zeta^t \in \{0,1\} \tag{4.39}$$

$$s_{i,\gamma^p+\zeta^p,\gamma^t+\zeta^t} \geq s_{j,\gamma^p,\gamma^t} + (\bar{p}_j + \zeta^p \hat{p}_j + \bar{t}_{ji} + \zeta^t \hat{t}_{ji}) x_{ji} + (\underline{T}_{id} - \bar{T}_{jd})(1 - x_{ji}) \qquad \forall i, j \in \mathcal{I}_{TSP}\,, \gamma^p \in [\Gamma^p]_0\,, \gamma^t \in [\Gamma^t]_0 \qquad \forall \zeta^p, \zeta^t \in \{0,1\} : \gamma^p + \zeta^p \leq \Gamma^p, \gamma^t + \zeta^t \leq \Gamma^t \tag{4.40}$$

$$s_{n+1} \geq s_{i,\Gamma^p-\zeta^p,\Gamma^t-\zeta^t} + (\bar{p}_i + \zeta^p \hat{p}_i + \bar{t}_{i,n+1} + \zeta^t \hat{t}_{i,n+1}) x_{i,n+1} \qquad \forall i \in \mathcal{I}_{TSP}\,, \zeta^t \in \{0,1\} \tag{4.41}$$

$$s_{n+1} \in [\underline{W}_{kd}, \bar{W}_{kd}] \tag{4.42}$$

$$\sum_{j \in \mathcal{I}_{TSP}} x_{ij} = 1 \qquad \forall i \in \mathcal{I}_{TSP} \tag{4.43}$$

$$\sum_{j \in \mathcal{I}_{TSP}} x_{ji} - \sum_{j \in \mathcal{I}_{TSP}} x_{ij} = 0 \qquad \forall i \in \mathcal{I}_{TSP} \tag{4.44}$$

$$\sum_{i \in \mathcal{I}_{TSP}} x_{ii} = 0 \qquad \forall i \in \mathcal{I}_{TSP} \tag{4.45}$$

$$\sum_{i \in \mathcal{I}_{TSP} \cup \{n+1\}} x_{0,i} = 1 \tag{4.46}$$

$$\sum_{i \in \mathcal{I}_{TSP} \cup \{0\}} x_{i,n+1} = 1 \tag{4.47}$$

$$x_{ij} \in \{0,1\} \qquad \forall i, j \in \mathcal{I}_{TSP} \cup \{0\} \cup \{n+1\} \tag{4.48}$$

$$s_{i,\gamma^p,\gamma^t}, s_{n+1} \geq 0 \qquad \forall i \in \mathcal{I}_{TSP}\,, \gamma^p \in [\Gamma^p]_0\,, \gamma^t \in [\Gamma^t]_0 \tag{4.49}$$

Objective Function (4.37) of this subproblem minimizes the total travel cost plus total wage cost. Constraints (4.38) ensure patients are visited within their time windows, even in a worst-case scenario. Constraints (4.39) set the first

visit's start time with respect to the caregiver's working start time and worst-case travel time. Constraints (4.40) ensure patients are visited after finishing the previous patient's visit and traveling, for all scenarios of service and travel times. Constraints (4.41) accordingly define the earliest return time to the depot. Constraint (4.42) ensures the caregiver returns within working hours. Constraints (4.43) guarantee that all patients are visited. Constraints (4.44) to (4.45) define flow conservation, and Constraints (4.46) to (4.47) specify the caregiver starts and ends at the depot.

The R-TSPTW subproblem has $\mathcal{O}\left((n_k)^2\right)$ binary variables, $\mathcal{O}\left(n_k \cdot \Gamma^p \cdot \Gamma^t\right)$ continuous variables, and $\mathcal{O}\left((n_k)^2 \cdot \Gamma^p \cdot \Gamma^t\right)$ constraints. Efficient solutions can be found using state-of-the-art MILP solvers, as caregivers in home healthcare typically visit at most $n_k = 8$ patients per day, and the uncertainty budgets w.l.o.g. satisfy $\Gamma^p \leq n_k$ and $\Gamma^t \leq n_k + 2$ (see Section 4.3.3.3). Furthermore, we emphasize that the objective function value does not include any of the dual variables. Hence, we globally store the R-TSPTW routes and their corresponding values, obviating the need to recalculate (4.37) – (4.49) for identical patient sets at other nodes of the B&B search tree.

Finally, if the R-TSPTW's objective function value exceeds the dual value u_d^*, the corresponding route yields an improving column; otherwise, we branch on the most promising patient with respect to the upper bound $\bar{v}_{kd}(i)$ in order to either find an improving assignment or prove that none exists. Thus, focusing on promising routes, our branch-and-bound procedure efficiently explores the solution space of daily routes.

4.4.3. Branching

At each level, the column generation procedure yields an optimal solution for the linear relaxation of the corresponding master problem, which may not be integer. To obtain an optimal integer solution, we integrate the column generation procedure at each level with a branch-and-bound tree, resulting in a nested branch-and-price method.

Upon termination of the column generation procedure with a solution λ, we check if the solution already constitutes an integer solution. If yes, we prune the node with this integer solution. Otherwise, we apply the following branching

Algorithm 2 Overview of first-level branching

for $i \in \mathcal{I}^{\mathcal{R}}$ *not yet branched upon* **do**
  $\mathrm{UB_{i,\emptyset}} \leftarrow 0$ **for** $k \in \mathcal{K}$ **do**
    $\mathrm{UB_{i,k}} \leftarrow \max_{\mathrm{D} \in \mathrm{day\_patterns_{ik}}} \mathrm{Val(D)}$
  $i\_\mathrm{branch} \leftarrow_{\mathrm{i}} \{\max_{\mathrm{k}} \{\mathrm{UB_{i,k}}\}\}$
return $i_$branch, $\mathrm{caregivers_{i\_branch,k}}$

strategy: In the first level, patients are branched upon caregivers. In the second level, patients are branched upon day patterns. At each level, we select the most promising remaining patient.

Specifically, for each patient i and each caregiver k, the set of feasible day patterns, denoted as $\mathrm{day\_patterns_{ik}}$, encompasses all subsets of days satisfying the correct number of visits per patient and the minimum number of days between any two visits. Optionally, it may ensure that visiting patient i on all days in the pattern does not exceed caregiver k's workload. The value of a day pattern is estimated using the upper bound

$$\begin{aligned} Val(D) := & \sum_{d \in D} R_{ki} \\ & - \sum_{d \in D} C^W_{k,d} \left(\bar{p}_i + \frac{1}{2} \left(\min_{j' \neq j'' \in \mathcal{I}^E_{k,d}} \bar{t}_{j',i} + \bar{t}_{i,j''} \right) \right) \\ & - \sum_{d \in D} \frac{1}{2} \left(\min_{j' \neq j'' \in \mathcal{I}^E_{k,d}} C^T_{d,j',i} + C^T_{d,i,j''} \right). \end{aligned} \tag{4.50}$$

The patient selected for branching maximizes the estimated value UB of the fixed assignment, as described in Procedure 2 and Procedure 3. Furthermore, the corresponding assignment of maximum value is selected for node selection.

Algorithm 3 Overview of second-level branching

for $i \in \mathcal{I}^{\mathcal{R}}_k$ *not yet branched upon* **do**
  $\mathrm{UB_{i,\emptyset}} \leftarrow 0$ **for** $D \in \mathrm{day\_patterns_{ik}}$ **do**
    $\mathrm{UB_{i,D}} \leftarrow \mathrm{Val(D)}$
  $i\_\mathrm{branch} \leftarrow_{\mathrm{i}} \{\max_{\mathrm{D}} \{\mathrm{UB_{i,D}}\}\}$
return $i_$branch, $\mathrm{day\_patterns_{i\_branch}}$

4.5. Computational Study

In this section, we evaluate the results obtained with our nested branch-and-price procedure using real-world data obtained from a prominent US healthcare agency based in Dallas, Texas [see Guo and Bard (2023))]. We address the agency's task of computing routes and schedules for the forthcoming week. At the end of each day of the current week, the agency decides whether to accept the patient requests received that day. The acceptance decisions are made simultaneously with assignment, routing, and scheduling decisions, while ensuring continuity of care for existing patients. We review the resulting weekly plan and derive management insights about different degrees of uncertainty and time windows.

4.5.1. Instances

In all, we solved 144 test instances generated from the agency's data set (see Appendix A.3.3.1 for details). These instances encompass 5,999 routine visits conducted by 36 caregivers during a four-week period from July 23 to August 20, 2018. The first week is set apart to establish an initial weekly plan using the greedy heuristic described in Section 4.4.2.1, while the subsequent three weeks are used to compute and evaluate optimal weekly home healthcare plans with the nested B&P procedure.

Section 4.5.1.1 outlines the fixed parameter values that remain constant across all instances. Section 4.5.1.2 introduces the factors—discipline, regional branch, time window size, and uncertainty budget—that we systematically vary using a full factorial design.

4.5.1.1. Fixed Factors

The parameters—revenue and cost, along with availability of caregivers and new patients—remain constant across all instances.

Revenue & Cost. The revenue per visit is calculated using the nursing hourly billing rate of 102 USD, as reported by the home healthcare agency under consideration. We assume that the services provided by registered nurses, occupational therapists, and speech therapists generate exactly this hourly revenue, while a

10% discount is applied to licensed practical nurses, skilled nurses, occupational therapist aides, and physical therapy assistants, and a 10% surcharge is applied for physical therapy. For hourly-paid caregivers, we refer to Table A.2 for wage cost. If caregivers are paid per visit, we subtract the wage cost, proportional to the expected service time, from the revenue of each visit. Salaried caregivers incur a fixed wage cost in this operational decision, which is the weekly number of work hours multiplied by the hourly wage cost. Travel cost is estimated at 0.56 USD per mile, based on the 2021 IRS standard mileage rate.

Caregivers. We establish a fixed group of caregivers across all weeks, determined based on the week of August 1, 2018. Caregivers are available from Monday to Friday, with full-time working hours from 9 am to 5 pm and part-time working hours from 8 am to 12 pm. The employment type and compensation details are presented in Table A.3 in Appendix A.3.3.2. All caregivers with a contract or working on an as-needed basis (PRN) are assumed to be available full-time. Each caregiver is assumed to be proficient in providing any service that represented at least 5% of the visits delivered by that caregiver in the historical data.

Patients. Patients are classified as either existing or new, depending on whether they were visited in the previous week. Existing patients retain their visits by the same caregiver on the same days. For new patients, however, the assignment needs to be determined. New patients are considered available every day of the week and compatible with any caregiver capable of providing the required service. The number of weekly visits required for each new patient is determined by their mean weekly visits in the historical data. The minimum number of days between routine visits is based on the minimum observed days in the data for each service. For all services this is 0 (but at most 1 visit per day), except for registered nursing, where it is 1 day.

4.5.1.2. Varied Factors

In order to evaluate the robustness of our home healthcare plans, we introduce uncertainty to each instance and then systematically vary the following factors in

a full factorial design: discipline, region, time window size, and uncertainty level. In the following, we provide a detailed description of each of these factors.

Disciplines. We explore two disciplines: nursing and physical therapy. Each discipline consists of specific services: Nursing includes skilled nursing (SN), registered nursing (RN), and licensed practical nursing (LPN) services. Physical therapy comprises physical therapy (PT) and physical therapy assistance (PTA). By default, a test instance represents one discipline, encompassing all routine visits within that discipline. The disciplines and services vary in terms of mean weekly visits per patient, average visit duration, and wage cost. An overview of these variations is provided in Table 4.1.

Regions. The two regional branches considered in our computational study differ in terms of population, patient count, average travel time and distance between any two patient locations. We named the branches based on the average travel time, computed using the OSRM routing engine (Luxen and Vetter 2011). Specifically, Region-T20 has an average travel time of 20 minutes (and 22.60 km) and Region-T60 has an average travel time of 60 minutes (and 74.69 km) between any two patient locations. A scatter plot depicting the patient locations in each region is shown in Figure 4.3. Region-T20 and Region-T60 are rural areas with a population below 100,000.

Time Windows. For existing patients, we explore three sizes of time windows: an exact appointment time (size: 0 minutes), a one-hour time window (size: 60 minutes), and a AM/PM/None time window categories based on the data set.

Table 4.1. Services in the data set

Service	Mean frequency (per patient and week)	Mean treatment period (per patient)	Expected duration	Wage (hourly)
SN	1 visit(s)	5 weeks	50 min	\$26.86
RN	1 visit(s)	2 weeks	55 min	\$42.80
LPN	1 visit(s)	4 weeks	50 min	\$26.86
PTA	2 visit(s)	5 weeks	50 min	\$31.01
PT	1 visit(s)	3 weeks	55 min	\$47.10

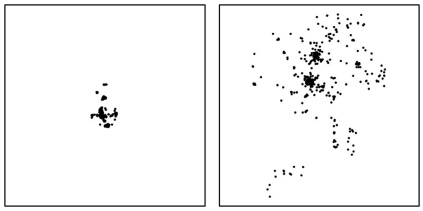

Figure 4.3. Scatter plots of patient locations in Region-T20 and Region-T60 (identical scaling, arranged in increasing order with respect to average travel time)

Uncertainty. We investigate four levels of uncertainty ranging from no uncertainty to anticipated deviations for all time estimates. We consider that patients and paths typically necessitate up to a 20% deviation from their expected service or travel time, respectively. We vary the uncertainty budgets, planning robustly against deviations of none, a quarter, half or all of the expected service and travel times. Table 4.2 summarizes these levels, characterized by the uncertainty budget Γ^p for service times, maximum service time deviations $\hat{p}$ (as a fraction of the expected service times $\bar{p}$), and the uncertainty budget Γ^t for travel times, alongside maximum travel time deviations $\hat{t}$ (as a fraction of the expected travel times $\bar{t}$).

Table 4.2. Levels of robustness

None	Low	Medium	High
$\Gamma^p = 0, \quad \Gamma^t = 0$	$\Gamma^p = 2, \quad \hat{p} = 20\%$	$\Gamma^p = 4, \quad \hat{p} = 20\%$	$\Gamma^p = 8, \quad \hat{p} = 20\%$
$\hat{p} = 20\%, \quad \hat{t} = 20\%$	$\Gamma^t = 2, \quad \hat{t} = 20\%$	$\Gamma^t = 4, \quad \hat{t} = 20\%$	$\Gamma^t = 8, \quad \hat{t} = 20\%$

4.5.2. Runtimes

All codes were written in Python and all computations performed in a Microsoft Windows 10 environment, running on an Intel Core i7-10510U 1.8 GHz processor with 16 GB memory operating in 64-bit mode. The second-level pricing problems were solved with Gurobi version 9.5.0. The implementation adheres to the specifications outlined in Section 4.4.

Table A.4 in Appendix A.3.3.2 displays the runtimes for the varied factors; i.e., region, discipline, uncertainty level, and time window size. The average runtime

is 591.01 seconds. Based on a correlation analysis, we can conclude that the following parameters primarily contribute to the runtime disparities.

Number of Patients and Caregivers. The factors impacting runtime most significantly are patient and caregiver counts, along with the daily number of new patient requests. This corresponds to the number of possible decisions, which increases exponentially with each of these factors. Table 4.3 summarizes the runtime effects.

Uncertainty Level. Higher robustness requires significantly longer runtimes, as shown in Table 4.4. Two main observations emerge: Firstly, there is a significant increase in runtimes up to a medium level of uncertainty, followed by a decrease. This pattern arises because setting the uncertainty budget equivalent to the number of daily patients (or paths) per caregiver mirrors the scenario of no uncertainty, where all data necessitates the maximum deviation (refer to Section 4.3.3.3). Therefore, runtimes peak when the uncertainty budget matches half the daily patients (and paths) per caregiver. Secondly, within the nested B&P procedure, uncertainty primarily impacts the second-level R-TSPTW pricing problem (see Problem 4.31). Since we used Gurobi to solve it, exploring faster solution methods for the R-TSPTW in future research appears promising to further improve runtimes under uncertainty (Bartolini et al. 2021, Lera-Romero et al. 2022).

Table 4.3. Runtimes by the total number n of considered patients, of which n_{new} patients are new (in seconds)

	$n_{new} = 15$ (3 per day)	$n_{new} = 25$ (5 per day)	
n = 1 - 24:	10.94	38.87	(K = 3.00)
n = 25 - 49:	1.23	179.34	(K = 4.90)
n = 50 - 74:	5.09	46.65	(K = 6.19)
n = 75 - 99:	75.70	158.62	(K = 6.42)
n = 100 - 124:	201.51	346.89	(K = 6.08)
n = 125 - 149:	-	2 922.67	(K = 6.00)
Average:	58.89	380.92	

Table 4.4. Runtimes by the number n of patients and the level of robustness (in seconds)

	None	Low	Medium	High
n = 1 - 24:	8.92	61.25	145.43	5.27
n = 25 - 49:	50.72	255.42	708.54	15.42
n = 50 - 74:	2.08	1.71	87.03	5.22
n = 75 - 99:	5.02	15.81	254.51	549.68
n = 100 - 124:	37.65	305.87	1 017.03	148.43
n = 125 - 149:	170.59	459.46	10 067.92	305.24
Average:	45.83	183.25	2 046.74	88.21

Time Windows. Time window size also impacts runtimes, albeit to a smaller extent compared to patient and caregiver counts. Tight time windows for existing patients (0 minutes) result in an average runtime of 34.17 seconds, narrow time windows (60 minutes) 42.27 seconds, and wide time windows (AM/PM/None) result in a runtime of 462.94 seconds. This decrease in runtimes with the time window size is primarily due to the number of feasible routes, which decreases significantly with shorter time windows.

Other. Finally, we identify factors with relevant but less significant effects on runtime based on the average correlation of all parameters when fixing the varied factors. As the week progresses and more capacity is used, the procedure accelerates (runtime correlation is -0.13 with the day index). A slight decrease in runtime occurs with a higher number v of weekly visits per patient, which we explain by the number $\binom{5}{v}$ of day patterns (correlation of -0.13). We also found that the optimality gap between the optimal solution and the first heuristic solution to have a negligible effect on runtime. Not surprisingly, these observations collectively suggest that faster runtimes are associated with fewer feasible solutions.

4.5.3. Managerial Insights

We investigate the impact of robustness and time window size on optimal weekly home healthcare plans, focusing on managerial insights. In Section 4.5.3.1, we analyze profit sensitivity, while Section 4.5.3.2 explores optimal acceptance, as-

signment, routing, and scheduling decisions. For a comprehensive overview of results across various regions, disciplines, time windows, and uncertainty levels, we refer to Table A.5 in Appendix A.3.3.2.

4.5.3.1. Profit Sensitivity

Across all instances, the weekly profit averages \$4,964.46, predominantly driven by a revenue of \$6,805.32, surpassing average weekly wage and travel costs of \$1,438.68 and \$402.18, respectively. The profitability of each visit, averaging \$96.97, emerges as the primary determinant of profit. Our subsequent analysis examines the effects of robustness and time window size on profit, revenue, wage cost, and travel cost, as outlined in Table 4.5 to Table 4.8.

Robustness. Figure 4.4 illustrates the profit, revenue, and costs across the different levels of robustness. We draw two key observations:

Firstly, in this case study, the almost constant disparity between revenue and profit curves indicates that robustness primarily decreases profit due to reduced patient acceptance. This is also evident from the observation that, on average, weekly plans with no uncertainty involve caregivers visiting on average 4.34 patients per day, while low, medium, and high uncertainty permit 4.01, 3.84, and 3.75 patients per day, respectively. This underscores the obvious yet crucial insight that fewer patient visits per day improve operational smoothness. However, it is also important to note that robustness can compromise health access if caregiver numbers cannot be augmented concurrently.

Secondly, the transition from no to low robustness decreases revenue by 5.05%, to medium robustness by 7.56%, and to high robustness by 14.20%. In particular,

Table 4.5. Profit by level of robustness and time window size
(in USD)

	None	Low	Medium	High
Wide:	5,624	5,361	5,095	4,880
Narrow:	5,297	5,026	4,825	4,767
Tight:	5,297	4,796	4,735	3,306
Average:	5,406	5,061	4,886	4,318

Table 4.6. Revenue by level of robustness and time window size
(in USD)

	None	Low	Medium	High
Wide:	7,496	7,201	6,916	6,709
Narrow:	7,151	6,860	6,661	6,598
Tight:	7,145	6,630	6,566	5,297
Average:	7,264	6,897	6,715	6,201

the decline in revenue is small between low and medium robustness. This finding suggests that in some cases, weekly plans can be robustified by adjusting schedules without compromising revenue, see Section 4.5.3.2.

Time Windows. On average, across all instances, wide time windows result in a profit of $5,412.46, narrow time windows in $4,922.69, and tight time windows

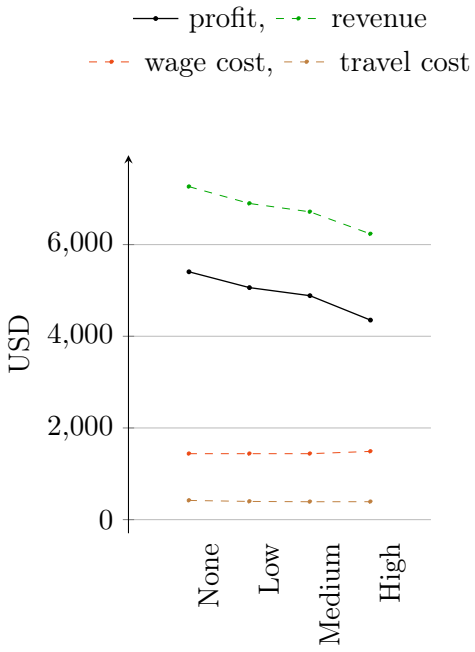

Figure 4.4. Levels of robustness

Table 4.7. Wage cost by level of robustness and time window size
(in USD)

	None	Low	Medium	High
Wide:	1,439	1,439	1,439	1,439
Narrow:	1,439	1,439	1,439	1,439
Tight:	1,439	1,439	1,439	1,599
Average:	1,439	1,439	1,439	1,492

in \$4,540.80. However, the impact of time windows is intricately linked with the level of robustness and warrants a detailed examination.

With no robustness, only a single scenario is considered, hence a single service time is determined. Wide time windows yield a profit of \$5,624.34, comprising a revenue of \$7,495.96, with wage and travel costs at \$1,438.68 and \$432.94, respectively. Narrow time windows induce more rejections, leading to a 4.60% decrease in revenue, and a 5.80% profit decline, respectively. Tight time windows only exacerbate this effect slightly, resulting in a 5.82% decrease in profit compared to wide time windows.

Introducing robustness leads to scenario-dependent service start times. Both the earliest and latest possible start times must fall within a patient's time window. Now, at a medium level of robustness, wide time windows yield a profit of \$5,095.61, a revenue of \$6,916.48, with wage and travel costs at \$1,438.68 and \$382.18, respectively. Once again, narrow time windows lead to more rejections, causing a comparable revenue decrease of 3.69%, and a profit decrease of 5.30%. However, under uncertainty, tight time windows exacerbate this trend further,

Table 4.8. Travel cost by level of robustness and time window size
(in USD)

	None	Low	Medium	High
Wide:	433	401	382	390
Narrow:	415	395	397	392
Tight:	409	395	392	392
Average:	419	397	390	391

resulting in a revenue decline of 5.07%, and a profit loss of 7.08% compared to wide time windows.

Overall, transitioning from wide time windows and no robustness to narrow time windows coupled with a medium level of robustness results in an average profit decrease of 14.20% in this case study. Primarily, this decline can be attributed to a reduced number of visits. However, the impact on profit fluctuates between different levels of robustness and time windows, suggesting a non-linear price of robustness and time window size.

4.5.3.2. Optimal Decisions

We now investigate optimal decisions for acceptance, assignment, routing, and scheduling. In contrast to many other approaches in the literature, our methodology makes these decisions simultaneously.

Acceptance. We first investigate the number of accepted patients as well as potential selection criteria across varying levels of robustness and time window sizes. Figure 4.5 illustrates patient acceptance and rejection for nursing in Region T-60.

If capacity permits, all patients are accepted, as the revenue from each patient typically exceeds the associated wage and travel costs. Selection only occurs when

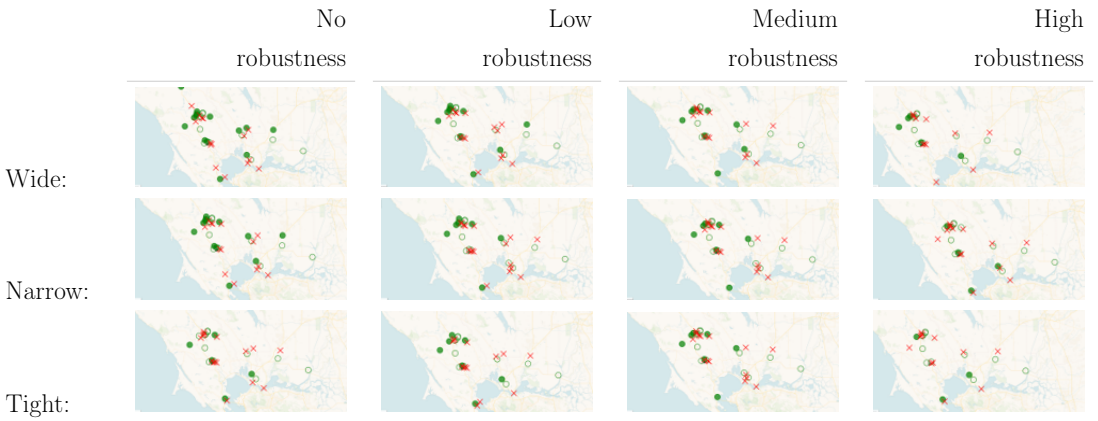

Figure 4.5. Acceptance decision by level of robustness and time window size (for Region T-60, Nursing)

capacity constraints necessitate rejecting patients. On average, 1.53 out of 5 new patients are accepted on day 1, 1.67 on day 2, 1.24 on day 3, 0.89 on day 4, and 0.60 on day 5, indicating a high capacity utilization for the home healthcare agency under consideration. As already indicated in terms of revenue changes, acceptance rates vary with different levels of robustness and time window sizes. For wide time windows, 6.27 patients are typically accepted, compared to 6.04 with narrow time windows, and 5.67 with tight time windows. Similarly, on average, 7.65, 6.17, 5.64, and 4.92 out of 25 new patients are accepted per week with no, low, medium, and high budget robustness, respectively. This represents a decrease in acceptance rates of 19.35%, 26.27%, and 35.69% for low, medium, and high robustness levels compared to no robustness.

Furthermore, the simple greedy heuristic from Section 4.4.2.1 demonstrates strong performance, achieving an average gap of 7.73% compared to the optimal solution, relative to the difference between the optimal solution value and the solution value when rejecting all patients, and averaged over all days where the optimal solution accepts at least one patient. Additionally, Figure 4.5 illustrates that new patients may be selected even if they live farther away than other candidates, indicating that geographic proximity is not necessarily critical. These observations suggest that a patient's contribution margin may take priority over geographical proximity in the decision-making process. Accordingly, we recommend selecting patients similar to the greedy heuristic, iteratively based on their potential value, determined by revenue minus an estimate of travel and wage costs.

Assignment. The decision of assigning new patients to caregivers and days carefully balances maximizing total revenue and minimizing total travel time and distance, while ensuring feasibility with respect to all patient time windows and work hours, even under uncertainty.

We find that geographic proximity holds significance primarily in settings with wide time windows. However, even in such cases, there is limited support for the traditional approach of pre-clustering caregivers by region. Instead, we observe geographic clusters by caregiver for each day independently, as depicted in Figure 4.6. To clarify this observation, let us consider two extreme cases theoretically: one where all patients require one visit per week and another where some

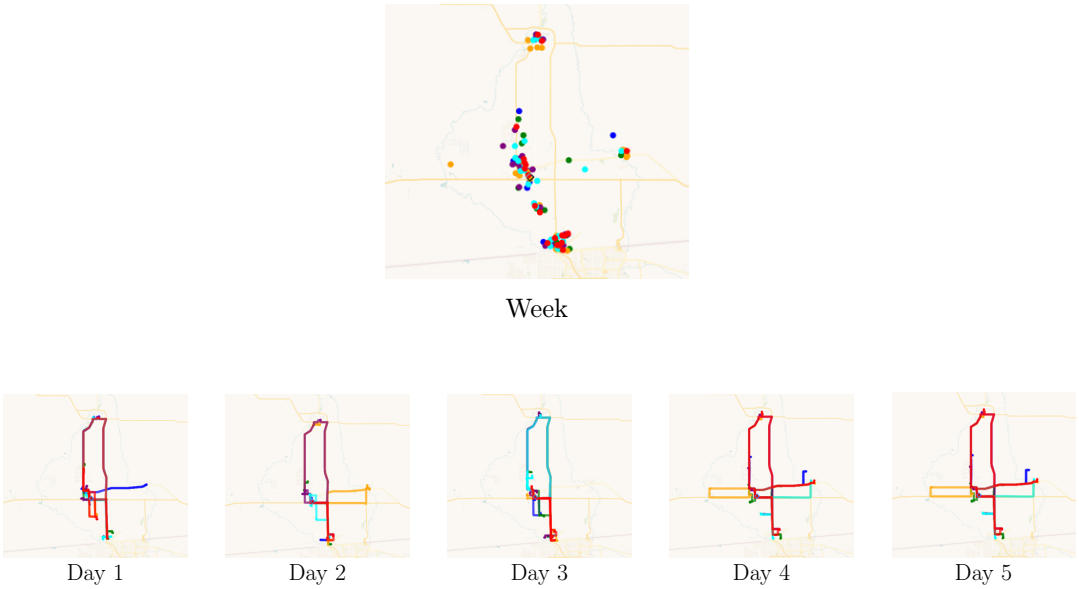

Figure 4.6. Illustration of patient-caregiver assignment by week and by day (for Region T-20, Nursing)

patients require visits on all days. In the former case, caregivers visit similar regions daily, aligning with pre-clustering regions by caregiver. Conversely, in the latter case, where patients need only one visit per week, freely assigning a new region for each caregiver each day may be more profitable. Thus, the conventional approach of pre-clustering caregivers by region seems advantageous only when time windows are wide and patients require multiple visits per week.

In settings with narrow or tight time windows, we cannot observe distinct patterns based on geographic proximity. We explain this with feasibility considerations. Even if nearby patients cannot be visited within their time windows, it remains more profitable to visit distant patients than none at all. As a result, geographical factors appear less influential with shorter time windows, further compounded by higher levels of robustness. In such situations, we discourage the use of clustering heuristics and suggest either computing optimal solutions or resorting to alternative heuristics like a simple greedy heuristic that iteratively fixes the most promising patient-caregiver-day patterns.

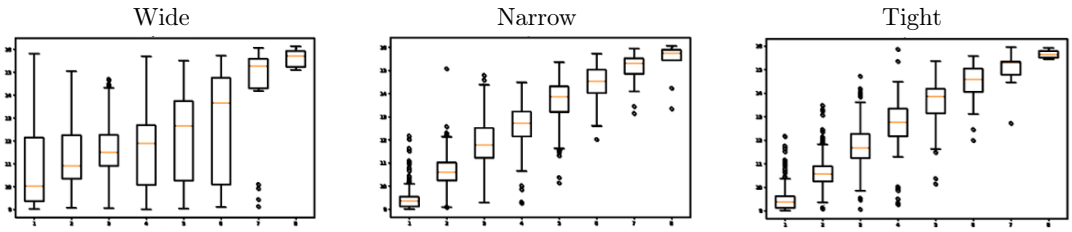

Figure 4.7. Daily schedules with medium robustness by time window size (for Region T-20, Nursing)

Routes and Schedules. Finally, we assess the daily routes and schedules across various levels of robustness and time window sizes. If fewer visits can be accommodated due to higher robustness or shorter time windows, the routes naturally adjust. However, if all visits can still be met, routes often require minimal or no adjustments to accommodate higher robustness or shorter time windows.

We find that routes may remain similar across different robustness levels if the set of visited patients remains unchanged. The main adjustments typically occur in the schedules, with discrepancies becoming more noticeable as the day progresses. While dependent on the estimated deviations of service and travel times, the worst-case start times change on average as follows. Compared to no robustness, the worst-case start time for the 8th visit is 29.30 minutes later for low robustness and 33.74 minutes later for medium robustness. With high robustness, there was no 8th visit on any day. Furthermore, as depicted in Figure 4.7, decreasing time window lengths tend to increase the scheduled time between any two visits. In summary, scheduling buffers between visits enables shorter time windows and increased robustness, with limited profit loss as long as revenue remains unaffected.

4.6. Conclusion

This work introduces a comprehensive decision tool for patient selection, caregiver assignment, and multi-day routing and scheduling, taking into account polyhedral uncertainty sets for travel and service times. It incorporates a variety of optimization criteria, including patient time windows, number of required visits,

caregiver availability, compatibility, continuity of care, and profit. The tool is designed to ensure optimal treatment for all selected patients, considering the factors influenced by operational planning. In particular, all plans are robust, ensuring timely availability for patients and caregivers even in worst-case scenarios, which can be adjusted through uncertainty budgets. At the same time, profit is maximized, calculated as revenue minus wage cost and travel cost.

We efficiently decompose the Robust Home Healthcare Routing and Scheduling Problem in several layers for a nested branch-and-price procedure. In the first layer, the problem is decomposed by caregiver, yielding weekly plans. Next, the problem of finding a weekly plan for each caregiver is decomposed by day, yielding daily routes for each. These routes are computed using a branch-and-bound procedure, with traveling salesman subproblems (R-TSPTW), and a multi-knapsack problem guiding the acceptance decision.

Our case study with real data from a large US home healthcare agency demonstrates that, while runtimes increase with shorter time windows and the level of robustness, all instances can be optimally solved within reasonable computation times. Surprisingly, geographic factors seem secondary to revenue and feasibility in the acceptance and assignment of new patients. Compared to parameter settings absent of robustness and time windows, introducing a medium level of robustness and one-hour time windows, and thus smoother operations and a higher service level, leads to an average profit loss of 14.20%. This is mainly due to decreased patient loads per caregiver per day. Nevertheless, some daily schedules can be robustified by scheduling more buffer times without affecting revenue.

Our work also points to several directions for future research. Firstly, the problem setting could be expanded. Alternative objectives such as preference satisfaction and workload balance, additional constraints for lunch breaks, caregiver dependencies, as well as uncertain caregiver absence are also relevant from a practical standpoint. Secondly, runtimes could be further accelerated. While we solved the R-TSPTW subproblems using Gurobi, it seems promising, particularly for settings with high uncertainty, to employ a problem-specific solution method. Thirdly, it would be worthwhile to investigate a variety of acceptance and assignment policies. This could include a comparison of our approach with advanced clustering approaches. More research is also needed on fair acceptance procedures for new patients.

5. Conclusion

5.1. Summary

In this dissertation, robust service levels are introduced for three key problems in operational healthcare management: appointment scheduling, predictive overbooking, and home healthcare routing and scheduling. These service levels guarantee adherence to individual waiting time guarantees for appointment scheduling and predictive overbooking, and adherence to time windows for home healthcare. They are designed robustly to withstand various uncertainty factors, ensuring acceptable outcomes even in worst-case scenarios. Robustness in the appointment scheduling framework considers service times within box uncertainty sets, while for predictive overbooking, it additionally accounts for uncertain no-show predictions within novel uncertainty sets for false positive and false negative predictions. In the home healthcare routing and scheduling problem, weekly schedules are robust to unfortunate service and travel times within polyhedral uncertainty sets, whose degree of conservatism can be adjusted using uncertainty budgets. Under these robust service levels, total cost is minimized.

All problems investigated in this dissertation are $\mathcal{NP}$-hard. For the appointment scheduling problem, this is proven by a reduction to the subset sum problem. Each problem is addressed using specific decomposition techniques. Analytical approaches are employed for the appointment scheduling problem and the predictive overbooking problem. These result in easy-to-implement optimal scheduling and sequencing rules for relevant special cases in the former, and the derivation of a mixed-integer linear program using total unimodularity and LP duality in the latter. The home care routing and scheduling problem, however, is notably more intricate due to its greater number of decisions and constraints, necessitating a numerical approach. To tackle this complexity, a nested branching and pricing approach is adopted. All problems are solved to optimality.

The proposed models and solution methods are tested using real-world data: the first two based on data from a radiology department at a large German university hospital and the last using data from a major US home healthcare agency. The computational results illustrate that robust waiting time guarantees are adhered to for nearly all test instances and ensure more equitably distributed waiting times for patients. Remarkably, in instances where any benchmark approach also provides acceptable patient waiting times, the introduction of waiting time guarantees may simultaneously lead to cost reductions. The increase in costs with higher levels of service and robustness is carefully evaluated for all problems. Overall, the computational findings underscore the potential of introducing worst-case service levels for individual patients to enhance equity in appointment scheduling and routing, while minimizing costs.

5.2. Key Managerial Insights

The findings in this dissertation provide the following key managerial insights into the implementation of robust service levels for appointment scheduling, predictive overbooking, and home healthcare routing and scheduling.

5.2.1. Robust Appointment Scheduling with Waiting Time Guarantees

In this problem, waiting time guarantees are introduced for all patients under uncertain service times given by box uncertainty sets, while minimizing worst-case costs incurred by idle time and overtime.

Complexity. The worst-case waiting times clearly arise in scenarios in which all patients require maximum service time. Moreover, worst-case cost scenarios have been identified, necessitating minimum service times initially (resulting in idle time) followed by maximum service times (leading to overtime) for all patients. Yet, even for identical waiting time guarantees for all patients, and costs $c_1 = \ldots = c_i > c_{i+1} = \ldots = c_{n+1} \geq 0$, $c_o = 0$, finding an optimal schedule that is robust with respect to these scenarios is $\mathcal{NP}$-hard.

Optimal Scheduling and Sequencing Rules. The following intuitive and easy-to-implement rules offer optimal solutions in relevant special cases. Assuming constant idle time costs, i.e., no timely preference for the occurrence of idle time, it is optimal to schedule patients in ascending order with $\bar{p}_j - \underline{p}_j + (1 + c_o)W_j$, i.e., the amount of uncertainty about their service times and their waiting tolerance. This rule also provides a deeper understanding of the optimality of the well-known Smallest Variance First rule and extends it to consider waiting time guarantees. Moreover, for the case of non-increasing idle time costs, including cases where earlier idle time may be valued more expensively than later idle time due to factors such as accumulated documentation work, optimality of the ASAP scheduling rule is proven. This rule sets each appointment time as early as possible while ensuring compliance with the waiting time guarantees.

Waiting Times, Idle Times, and Overtime with Real-World Data. A case study conducted using data from the radiology department of a large German university hospital demonstrates the effectiveness of robust appointment scheduling with waiting time guarantees. Specifically, it shows that more than 96% of appointments adhere to a 30-minute waiting time guarantee, effectively preventing excessive waiting times. Moreover, in comparison to weighted sum approaches, the proposed approach offers greater precision in controlling waiting times, but also idle times and overtime. Moreover, the introduction of waiting time guarantees results is conducted with a minimal increase in worst-case costs. Generally, in environments where most patients experience acceptable waiting times, implementing RASWTG provides both waiting time guarantees and cost reduction. This cost reduction typically accompanies a slight increase in average waiting times. Conversely, in environments where many patients experience prolonged waiting times, RASWTG effectively prevents excessive waiting times with a minimal increase in total worst-case cost.

5.2.2. Robust Predictive Overbooking with Waiting Time Guarantees

In this problem, waiting time guarantees are introduced for all patients, considering not only uncertain service times but also individual no-show predictions. Robustness against false predictions is managed through uncertainty budgets, which anticipate both false positives and false negatives. The objective remains to minimize the worst-case cost associated with idle time and overtime.

Decomposability. To derive a mixed-integer linear program, the problem can be decomposed as follows. Initially, the worst-case waiting time constraint is broken down on a per-patient basis, allowing for its linearization through total unimodularity and LP duality. Secondly, the objective function can again be decomposed into two distinct components: an earlier idle time segment and a subsequent overtime incurring segment. Idle time is further divided into blocks.

Waiting Times, Idle Times, and Overtime. Our computational analysis reveals that a state-of-the-art MILP solver efficiently finds optimal solutions for instances involving up to eight patients within reasonable runtimes. These results consistently demonstrate adherence to a 30-minute waiting time guarantee in over 96% of cases across various scenarios. We evaluate the impact of prediction uncertainty, noting a decrease in total idle time by up to 8% and a decrease by up to 24% in overtime with perfect predictions compared to no individual predictions. This enhancement is even more significant for the completion time of the last appointment. Interestingly, we observe a slightly more adverse effect on total cost from false positives compared to false negatives. Moreover, surprisingly, a higher no-show rate decreases waiting times and simultaneously enables a cost reduction through our overbooking approach. This suggests that the proposed overbooking approach is capable of effectively mitigating the negative effects implied by no-shows.

5.2.3. Robust Home Healthcare Operations

This problem entails simultaneous decisions regarding acceptance, patient-caregiver assignment, routing, and scheduling over a multi-day planning horizon, involving

multiple caregivers. All health-related criteria are enforced as hard constraints, along with all cost-related criteria in the objective function.

Exact Solution Method. The complexity of the problem can be addressed by decomposing the creation of weekly plans, initially by caregiver and then by day. Subsequently, for each day, a Knapsack problem can be employed to identify promising assignments, which ultimately breaks the routing and scheduling decision down to a small robust traveling salesman subproblem with time windows. Leveraging this decomposable problem structure, a nested branch-and-price procedure has been proposed. Our case study, based on real data from a large US home healthcare agency, demonstrates that while runtimes increase with shorter time windows and higher levels of robustness, all test instances can be optimally solved within reasonable computation times.

The Price of Robust Service Levels. Robust service levels entail robust adherence to time windows. Introducing a medium level of robustness along with one-hour time windows, in contrast to parameter settings devoid of robustness and time windows, results in smoother operations and a higher service level, albeit with an average profit loss of 14.20%. This decrease is primarily attributed to reduced patient loads per caregiver and day. Interestingly, geographic considerations seem to be of secondary importance in the acceptance and assignment of new patients, with a simple greedy heuristic proving effective instead. Furthermore, certain daily schedules can be enhanced with additional buffer times without affecting revenue.

5.3. Limitations and Recommendations

This dissertation demonstrates the potential of robust service levels, particularly waiting time guarantees, in improving equity and reliability of waiting times, offering a straightforward and customizable framework. It requires further research across diverse healthcare settings to identify patient-specific thresholds for acceptable waiting times, considering both health outcomes and patient satisfaction. Moreover, it is important to note that robust waiting time guarantees may result in increased average waiting times compared to traditional methods, and

may lead to reduced capacity utilization in settings with excessive waiting times. Additionally, it is important to conduct further evaluation of the cost-effectiveness of waiting time guarantees in other healthcare settings and to compare the performance with alternative fairness approaches, such as those proposed by Qi (2017), Samorani et al. (2021), and Benjaafar et al. (2023).

Moreover, the chosen uncertainty sets are inherently pivotal for ensuring the robustness and performance of the models. While the existing box uncertainty sets for appointment scheduling already result in $\mathcal{NP}$-hardness, and in reasonable performance for instances with up to ten patients, they seem excessively conservative when handling larger instances. Thus, there is a need to consider uncertainty budgets to control the level of conservatism. To this end, as introducing uncertainty budgets for service times further increases the complexity of the appointment scheduling problem, it might be advisable to focus on the simpler and practically applicable case of constant idle time costs.

Furthermore, it is essential to find more efficient solution methods for the predictive overbooking problem. Moreover, the nested branch-and-price method for the home healthcare problem could be further accelerated using a sophisticated solution method for the robust traveling salesman subproblems with time windows. It is important to note that the proposed decomposition approach for addressing the home healthcare routing and scheduling problem relies on the assumption of caregivers' independence, warranting further exploration of settings involving interdependent caregivers. Clearly, the efficiency of all approaches should also be validated with other data and in different contexts.

Finally, the trade-off between waiting times, idle times, and overtime should be discussed with ethicists and practitioners. For instance, a combination of robust service levels with optimizing expected capacity utilization appears intriguing. Additionally, the acceptance decision deserves more attention and the outcomes of predictive overbooking necessitate further consideration from an ethical standpoint. All in all, appointment scheduling problems should be further examined within a broader framework of operational, tactical, and strategic decisions to enhance healthcare delivery.

5.4. Future Research Directions

This dissertation points towards several avenues for future research, ranging from further examination of problem-specific insights to exploring the transferability of the applied approaches. The following areas are highlighted for further investigation:

Appointment Scheduling and Overbooking. Firstly, it appears valuable to incorporate uncertainty budgets controlling conservatism with respect to service time uncertainty. Developing optimal scheduling and sequencing rules for this extended problem presents an intriguing avenue for exploration. Secondly, integrating robust waiting time guarantees into more comprehensive appointment scheduling problems and decisions, especially incorporating patient acceptance or rejection decisions, seems promising. Thirdly, more efficient solution methods for larger instance sizes are required, in particular for robust predictive overbooking.

Home Healthcare Routing and Scheduling. In addition to the presented approach, practitioners have explicitly expressed interest in robustness with respect to uncertain caregiver absence. Moreover, alternative objectives such as preference satisfaction and workload balance, additional constraints for lunch breaks, and caregiver dependencies are also relevant from a practical standpoint. Apart from expanding the problem scope, investigating optimal acceptance and assignment policies, with a focus on fairness, would be worthwhile.

Generalizable Concepts. Firstly, there is promise in further exploring the proposed uncertainty budgets for false predictions. Secondly, the nested branch-and-price procedure can be applied to other related routing and scheduling problems. Thirdly, the potential for applying and customizing the proposed robust service levels extends to a wider range of problems in healthcare and beyond.

In general, and concerning the application of the proposed robust service levels to other problems, more research is required to further grasp the consequences of operational decisions and, in turn, to refine optimization models. Fostering collaboration with ethicists and medical experts is paramount to developing responsible optimization models.

Bibliography

Agra A, Christiansen M, Figueiredo R, Hvattum LM, Poss M, Requejo C (2013) The robust vehicle routing problem with time windows. *Computers & Operations Research* 40(3):856–866.

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

Aiken LH, Clarke SP, Sloane DM, Sochalski J, Silber JH (2002) Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *Journal of the American Medical Association* 288(16):1987–1993.

Alaeddini A, Yang K, Reddy C, Yu S (2011) A probabilistic model for predicting the probability of no-show in hospital appointments. *Health Care Management Science* 14(2):146–157.

Alkaabneh F, Diabat A (2023) A multi-objective home healthcare delivery model and its solution using a branch-and-price algorithm and a two-stage meta-heuristic algorithm. *Transportation Research Part C: Emerging Technologies* 147:103838.

Ayer T, Keskinocak P, Swann J (2014) Research in public health for efficient, effective, and equitable outcomes. *Bridging Data and Decisions*, 216–239 (INFORMS).

Bae SH, Fabry D (2014) Assessing the relationships between nurse work hours/overtime and nurse and patient outcomes: systematic literature review. *Nursing Outlook* 62(2):138–156.

Bailey NT (1952) A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. *Journal of the Royal Statistical Society: Series B (Methodological)* 14(2):185–199.

Bard JF, Shao Y, Qi X, Jarrah AI (2014) The traveling therapist scheduling problem. *IIE Transactions* 46(7):683–706.

Bartolini E, Goeke D, Schneider M, Ye M (2021) The robust traveling salesman problem with time windows under knapsack-constrained travel time uncertainty. *Transportation Science* 55(2):371–394.

Batool T, Abuelnoor M, El Boutari O, Aloul F, Sagahyroon A (2021) Predicting hospital no-shows using machine learning. *2020 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, 142–148 (IEEE).

Bean AG, Talaga J (1992) Appointment breaking: causes and solutions. *Journal of Health Care Marketing* 12(4).

Ben-Tal A, El Ghaoui L, Nemirovski A (2009) *Robust optimization*, volume 28 (Princeton university press).

Ben-Tal A, Nemirovski A (1998) Robust convex optimization. *Mathematics of Operations Research* 23(4):769–805.

Ben-Tal A, Nemirovski A (2000) Robust solutions of linear programming problems contaminated with uncertain data. *Mathematical Programming* 88:411–424.

Bendavid I, Marmor YN, Shnits B (2018) Developing an optimal appointment scheduling for systems with rigid standby time under pre-determined quality of service. *Flexible Services and Manufacturing Journal* 30(1):54–77.

Benjaafar S, Chen D, Wang R, Yan Z (2023) Appointment scheduling under a service-level constraint. *Manufacturing & Service Operations Management* 25(1):70–87.

Bertsimas D, Farias VF, Trichakis N (2011) The price of fairness. *Operations Research* 59(1):17–31.

Bertsimas D, Sim M (2004) The price of robustness. *Operations Research* 52(1):35–53.

Bielen F, Demoulin N (2007) Waiting time influence on the satisfaction-loyalty relationship in services. *Managing Service Quality: An International Journal* 17(2):174–193.

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

Bosch PMV (1997) Scheduling and sequencing arrivals to a stochastic service system. Doctoral Dissertation, Air Force Institute of Technology. ProQuest Dissertations and Theses Database (AAT 9813076).

Bouch A, Kuchinsky A, Bhatti N (2000) Quality is in the eye of the beholder: Meeting users' requirements for internet quality of service. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems 2000*, 297–304.

Cabana MD, Jee SH, et al. (2004) Does continuity of care improve patient outcomes. *The Journal of Family Practice* 53(12):974–980.

Caffrey C, Sengupta M, Moss A, Harris-Kojetin L, Valverde R (2011) Home health care

and discharged hospice care patients. *National Health Statistics Reports* (38):1–27, ISSN 2164-8344.

Camacho F, Anderson R, Safrit A, Jones AS, Hoffmann P (2006) The relationship between patient's perceived waiting time and office-based practice satisfaction. *North Carolina Medical Journal* 67(6):409–413.

Carello G, Lanzarone E (2014) A cardinality-constrained robust model for the assignment problem in home care services. *European Journal of Operational Research* 236(2):748–762.

Carello G, Lanzarone E, Laricini D, Servilio M (2017) Handling time-related demands in the home care nurse-to-patient assignment problem with the implementor-adversarial approach. *International Conference on Health Care Systems Engineering*, 87–97 (Springer).

Carreras-García D, Delgado-Gómez D, Llorente-Fernández F, Arribas-Gil A (2020) Patient no-show prediction: a systematic literature review. *Entropy* 22(6):675.

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

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

Cayirli T, Yang KK, Quek SA (2012) A universal appointment rule in the presence of no-shows and walk-ins. *Production and Operations Management* 21(4):682–697.

Centers for Medicare & Medicaid Services N (2020) National health expenditures accounts. URL http://dx.doi.org/https://www.cdc.gov/nchs/data/hus/2020-2021/HExpType.pdf.

Chan KS, Wan EYF, Chin WY, Cheng WHG, Ho MK, Yu EYT, Lam CLK (2021) Effects of continuity of care on health outcomes among patients with diabetes mellitus and/or hypertension: a systematic review. *BMC Family Practice* 22(1):1–13.

Chen J, Goldstein IH, Lin WC, Chiang MF, Hribar MR (2020) Application of machine learning to predict patient no-shows in an academic pediatric ophthalmology clinic. *AMIA Annual Symposium Proceedings*, volume 2020, 293 (American Medical Informatics Association).

Cissé M, Yalçındağ S, Kergosien Y, Şahin E, Lenté C, Matta A (2017) OR problems related to home health care: A review of relevant routing and scheduling problems. *Operations Research for Health Care* 13:1–22.

CMS (2020) Quality of patient care star ratings methodology. URL https://www.medicare.gov/care-compare/resources/home-health/quality-of-patient-care.

Dai T, Tayur S (2020) OM forum—healthcare operations management: a snapshot of emerging research. *Manufacturing & Service Operations Management* 22(5):869–887.

Dantas LF, Fleck JL, Oliveira FLC, Hamacher S (2018) No-shows in appointment scheduling–a systematic literature review. *Health Policy* 122(4):412–421.

Davis MM, Heineke J (1998) How disconfirmation, perception and actual waiting times impact customer satisfaction. *International Journal of Service Industry Management* 9(1):64–73.

de Kemp MA, Mandjes M, Olver N (2021) Performance of the smallest-variance-first rule in appointment sequencing. *Operations Research* 69(6):1909–1935.

Deb K, Sindhya K, Hakanen J (2016) Multi-objective optimization. *Decision sciences*, 161–200 (CRC Press).

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

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

Desrosiers J, Lübbecke ME (2011) Branch-price-and-cut algorithms. *Encyclopedia of Operations Research and Management Science. John Wiley & Sons, Chichester* 109–131.

Deyo RA, Inui TS (1980) Dropouts and broken appointments: a literature review and agenda for future research. *Medical Care* 18(11):1146–1157.

Di Mascolo M, Martinez C, Espinouse ML (2021) Routing and scheduling in home health care: A literature survey and bibliometric analysis. *Computers & Industrial Engineering* 158:107255.

Dohn A, Mason A (2013) Branch-and-price for staff rostering: An efficient implementation using generic programming and nested column generation. *European Journal of Operational Research* 230(1):157–169.

El Ghaoui L, Lebret H (1997) Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications* 18(4):1035–1064.

El Ghaoui L, Oustry F, Lebret H (1998) Robust solutions to uncertain semidefinite programs. *SIAM Journal on Optimization* 9(1):33–52.

Fan G, Deng Z, Ye Q, Wang B (2021) Machine learning-based prediction models for patients no-show in online outpatient appointments. *Data Science and Management* 2:45–52.

Fathollahi-Fard AM, Ahmadi A, Goodarzian F, Cheikhrouhou N (2020a) A bi-objective home healthcare routing and scheduling problem considering patients' satisfaction in a fuzzy environment. *Applied Soft Computing* 93:106385.

Fathollahi-Fard AM, Ahmadi A, Karimi B (2020b) A robust optimization for a home healthcare routing and scheduling problem considering greenhouse gas emissions and stochastic travel and service times. *Green Transportation and New Advances in Vehicle Routing Problems*, 43–73 (Springer).

Fathollahi-Fard AM, Ahmadi A, Karimi B (2021) Multi-objective optimization of home healthcare with working-time balancing and care continuity. *Sustainability* 13(22):12431.

Fathollahi-Fard AM, Ahmadi A, Karimi B (2022) Sustainable and robust home healthcare logistics: A response to the covid-19 pandemic. *Symmetry* 14(2):193.

Fikar C, Hirsch P (2017) Home health care routing and scheduling: A review. *Computers & Operations Research* 77:86–95.

Fries BE (1976) Bibliography of operations research in health-care systems. *Operations Research* 24(5):801–814.

Garey MR, Johnson DS (1979) *Computers and intractability: A guide to the theory of NP-hardness* (San Fransisco: WH Freeman).

Garuda SR, Javalgi RG, Talluri VS (1998) Tackling no-show behavior: a market-driven approach. *Health Marketing Quarterly* 15(4):25–44.

Gier J (2017) Missed appointments cost the US healthcare system 150 USD each year. *Health Management Technology* 2, URL https://www.hcinnovationgroup.com/clinical-it/article/13008175/missed-appointments-cost-the-us-healthcare-system-150b-each-year.

Grand View Research (2021) Home healthcare market size & share report. URL http://dx.doi.org/https://www.grandviewresearch.com/industry-analysis/home-healthcare-industry/request/rs15#.

Grenouilleau F, Lahrichi N, Rousseau LM (2020) New decomposition methods for home care scheduling with predefined visits. *Computers & Operations Research* 115:104855.

Guo J, Bard JF (2023) A three-step optimization-based algorithm for home healthcare delivery. *Socio-Economic Planning Sciences* 101517.

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

Heching A, Hooker JN (2016) Logic-based Benders decomposition for multiagent scheduling with sequence-dependent costs. *ISAIM*, URL https://isaim2016.cs.ou.edu/papers/ISAIM2016_CPOR_Heching_Hooker.pdf.

Heching A, Hooker JN, Kimura R (2019) A logic-based Benders approach to home healthcare delivery. *Transportation Science* 53(2):510–522.

Hennig F, Nygreen B, Lübbecke ME (2012) Nested column generation applied to the crude oil tanker routing and scheduling problem with split pickup and split delivery. *Naval Research Logistics* 59(3-4):298–310.

Hermida RC, Ayala DE, Smolensky MH, Fernandez JR, Mojon A, Portaluppi F (2016) Chronotherapy with conventional blood pressure medications improves management of hypertension and reduces cardiovascular and stroke risks. *Hypertension Research* 39(5):277–292.

Hill CJ, Joonas K (2005) The impact of unacceptable wait time on health care patients' attitudes and actions. *Health Marketing Quarterly* 23(2):69–87.

Hirello L, Pulok MH, Hajizadeh M (2022) Equity in healthcare utilization in Canada's publicly funded health system: 2000–2014. *The European Journal of Health Economics* 23(9):1519–1533.

Huang XM (1994) Patient attitude towards waiting in an outpatient clinic and its applications. *Health Services Management Research* 7(1):2–8.

Issabakhsh M, Lee S, Kang H (2021) Scheduling patient appointment in an infusion center: a mixed integer robust optimization approach. *Health Care Management Science* 24(1):117–139.

Jafarnia-Jahromi M, Jain R (2020) Non-indexability of the stochastic appointment scheduling problem. *Automatica* 118:109016.

Jiang R, Ryu M, Xu G (2019) Data-driven distributionally robust appointment scheduling over Wasserstein balls. Preprint, submitted July 7, https://arxiv.org/abs/1907.03219.

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

Kassenärztliche Bundesvereinigung (2021) Wartezeit in Arztpraxen in Deutschland in den Jahren 2008 bis 2021. Statista GmbH. Accessed: March

31, 2024. https://de.statista.com/statistik/daten/studie/706887/umfrage/umfrage-zur-wartezeiten-in-arztpraxen-in-deutschland/.

Kc DS, Scholtes S, Terwiesch C (2020) Empirical research in healthcare operations: Past research, present understanding, and future opportunities. *Manufacturing & Service Operations Management* 22(1):73–83.

Keskinocak P, Savva N (2020) A review of the healthcare-management (modeling) literature published in Manufacturing & Service Operations Management. *Manufacturing & Service Operations Management* 22(1):59–72.

Kong Q, Lee CY, Teo CP, Zheng Z (2013) Scheduling arrivals to a stochastic service delivery system using copositive cones. *Operations Research* 61(3):711–726.

Kong Q, Lee CY, Teo CP, Zheng Z (2016) Appointment sequencing: Why the smallest-variance-first rule may not be optimal. *European Journal of Operational Research* 255(3):809–821.

Kong Q, Li S, Liu N, Teo CP, Yan Z (2020) Appointment scheduling under time-dependent patient no-show behavior. *Management Science* 66(8):3480–3500.

Krischer JP (1980) An annotated bibliography of decision analytic applications to health care. *Operations Research* 28(1):97–113.

Kumar P, Kalwani MU, Dada M (1997) The impact of waiting time guarantees on customers' waiting experiences. *Marketing Science* 16(4):295–314.

Kuo YH, Balasubramanian H, Chen Y (2020) Medical appointment overbooking and optimal scheduling: tradeoffs between schedule efficiency and accessibility to service. *Flexible Services and Manufacturing Journal* 32(1):72–101.

LaGanga LR, Lawrence SR (2007) Clinic overbooking to improve patient access and increase provider productivity. *Decision Sciences* 38(2):251–276.

LaGanga LR, Lawrence SR (2012) Appointment overbooking in health care clinics to improve patient service and clinic performance. *Production and Operations Management* 21(5):874–888.

Lampert MT (2021) Gesundheitliche Ungleichheit. Accessed: March 31, 2024. https://www.bpb.de/kurz-knapp/zahlen-und-fakten/datenreport-2021/gesundheit/330144/zusammenfassung/.

Lee C, Lee K, Park S (2012) Robust vehicle routing problem with deadlines and travel time/demand uncertainty. *Journal of the Operational Research Society* 63(9):1294–1306.

Lera-Romero G, Miranda Bront JJ, Soulignac FJ (2022) Dynamic programming for

the time-dependent traveling salesman problem with time windows. *INFORMS Journal on Computing* 34(6):3292–3308.

Li Y, Tang SY, Johnson J, Lubarsky DA (2019) Individualized no-show predictions: Effect on clinic overbooking and appointment reminders. *Production and Operations Management* 28(8):2068–2086.

Lindley DV (1952) The theory of queues with a single server. *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 48, 277–289 (Cambridge University Press).

Liu D, Shin WY, Sprecher E, Conroy K, Santiago O, Wachtel G, Santillana M (2022) Machine learning approaches to predicting no-shows in pediatric medical appointment. *NPJ Digital Medicine* 5(1):50.

Liu R, Yuan B, Jiang Z (2017) Mathematical model and exact algorithm for the home care worker scheduling and routing problem with lunch break requirements. *International Journal of Production Research* 55(2):558–575.

Liu R, Yuan B, Jiang Z (2019) A branch-and-price algorithm for the home-caregiver scheduling and routing problem with stochastic travel and service times. *Flexible Services and Manufacturing Journal* 31(4):989–1011.

Lu M, Shen ZJM (2021) A review of robust operations management under model uncertainty. *Production and Operations Management* 30(6):1927–1943.

Lübbecke ME, Desrosiers J (2005) Selected topics in column generation. *Operations Research* 53(6):1007–1023.

Luo J, Kulkarni VG, Ziya S (2012) Appointment scheduling under patient no-shows and service interruptions. *Manufacturing & Service Operations Management* 14(4):670–684.

Luxen D, Vetter C (2011) Real-time routing with openstreetmap data. *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 513–516, GIS '11 (New York, NY, USA: ACM), ISBN 978-1-4503-1031-4, URL http://dx.doi.org/10.1145/2093973.2094062.

Mak HY, Rong Y, Zhang J (2015) Appointment scheduling with limited distributional information. *Management Science* 61(2):316–334.

Makboul S, Kharraja S, Abbassi A, El Hilali Alaoui A (2021) A robust home health care scheduling and routing approach with time windows and synchronization constraints under travel time and service time uncertainty. *IFIP International Conference on Advances in Production Management Systems*, 391–402 (Springer).

Marburger Bund (2013) Klinikärzte: Wie hoch ist Ihre tatsächliche Wochenarbeitszeit inkl. Überstunden und Bereitschaftsdienste im Durchschnitt? Statista GmbH. Accessed: March 31, 2024. https://de.statista.com/statistik/daten/studie/257383/umfrage/umfrage-zu-den-arbeitszeiten-von-klinikaerzten-in-deutschland/.

Marynissen J, Demeulemeester E (2019) Literature review on multi-appointment scheduling problems in hospitals. *European Journal of Operational Research* 272(2):407–419.

McKinsey (2022) From facility to home: How healthcare could shift by 2025. URL http://dx.doi.org/https://www.mckinsey.com/industries/healthcare/our-insights/from-facility-to-home-how-healthcare-could-shift-by-2025.

Mittal S, Schulz AS, Stiller S (2014) Robust appointment scheduling. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2014)*, volume 28 of *Leibniz International Proceedings in Informatics*, 356–370.

Mohammadi I, Wu H, Turkcan A, Toscos T, Doebbeling BN (2018) Data analytics and modeling for appointment no-show in community health centers. *Journal of Primary Care & Community Health* 9:2150132718811692.

Morris S, Sutton M, Gravelle H (2005) Inequity and inequality in the use of health care in england: an empirical investigation. *Social science & Medicine* 60(6):1251–1266.

Moschis GP, Bellenger DN, Curasi CF (2003) What influences the mature consumer? *Marketing Health Services* 23(4):16–21.

Munari P, Moreno A, De La Vega J, Alem D, Gondzio J, Morabito R (2019) The robust vehicle routing problem with time windows: compact formulation and branch-price-and-cut method. *Transportation Science* 53(4):1043–1066.

Naderi B, Begen MA, Zaric GS, Roshanaei V (2023) A novel and efficient exact technique for integrated staffing, assignment, routing, and scheduling of home care services under uncertainty. *Omega* 116:102805.

Naji W, Masmoudi M, Mellouli R (2017) A robust-MILP for synchronized-mTSPTW: Application to home health care under uncertainties. *2017 4th International Conference on Control, Decision and Information Technologies*, 1089–1094 (IEEE).

NHS (2019) Missed GP appointments costing NHS millions. Accessed: March 06, 2024. https://www.england.nhs.uk/2019/01/missed-gp-appointments-costing-nhsmillions/2019/.

OECD (2020) Waiting times for health services: next in line. *OECD Health Policy*

Studies Accessed: March 31, 2024. https://stats.oecd.org/Index.aspx?QueryId=49344.

of Labor Statistics B (2022) *Occupational outlook handbook* (US Department of Labor and Bureau of Statistics Staff and United States), URL https://www.bls.gov/ooh/home.htm.

Osakwe ZT, Larson E, Andrews H, Shang J (2019) Activities of daily living of home health care patients. *Home Healthcare Now* 37(3):165.

Pecin D, Contardo C, Desaulniers G, Uchoa E (2017) New enhancements for the exact solution of the vehicle routing problem with time windows. *INFORMS Journal on Computing* 29(3):489–502.

Porter ME, et al. (2009) A strategy for health care reform—toward a value-based system. *N Engl J Med* 361(2):109–112.

Qi J (2017) Mitigating delays and unfairness in appointment systems. *Management Science* 63(2):566–583.

Qiu H, Wang D, Yin Y, Cheng TE, Wang Y (2022) An exact solution method for home health care scheduling with synchronized services. *Naval Research Logistics* 69(5):715–733.

Rahimian H, Mehrotra S (2022) Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization* 3:4.

Rasmussen MS, Justesen T, Dohn A, Larsen J (2012) The home care crew scheduling problem: Preference-based visit clustering and temporal dependencies. *European Journal of Operational Research* 219(3):598–610.

Robinson LW, Chen RR (2003) Scheduling doctors' appointments: optimal and empirically-based heuristic policies. *IIE Transactions* 35(3):295–307.

Salah H, Srinivas S (2022) Predict, then schedule: Prescriptive analytics approach for machine learning-enabled sequential clinical scheduling. *Computers & Industrial Engineering* 169:108270.

Salazar LHA, Parreira WD, Fernandes AMdR, Leithardt VRQ (2022) No-show in medical appointments with machine learning techniques: A systematic literature review. *Information* 13(11):507.

Samorani M, Harris S (2019) The impact of probabilistic classifiers on appointment scheduling with no-shows. *Fortieth International Conference on Information Systems, Munich*, URL http://dx.doi.org/10.2139/ssrn.3463501.

Samorani M, Harris SL, Blount LG, Lu H, Santoro MA (2021) Overbooked and over-

looked: machine learning and racial bias in medical appointment scheduling. *Manufacturing & Service Operations Management* 24(6):2825–2842.

Samorani M, LaGanga LR (2015) Outpatient appointment scheduling given individual day-dependent no-show predictions. *European Journal of Operational Research* 240(1):245–257.

Sang P, Begen MA, Cao J (2021) Appointment scheduling with a quantile objective. *Computers & Operations Research* 132:105295.

Schulz AS, Udwani R (2019) Robust appointment scheduling with heterogeneous costs. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2014)*, volume 145 of *Leibniz International Proceedings in Informatics*, 25:1–25:17.

Shahnejat-Bushehri S, Tavakkoli-Moghaddam R, Boronoos M, Ghasemkhani A (2021) A robust home health care routing-scheduling problem with temporal dependencies under uncertainty. *Expert Systems with Applications* 182:115209.

Shapiro A, Dentcheva D, Ruszczynski A (2021) *Lectures on stochastic programming: modeling and theory* (Society for Industrial and Applied Mathematics).

Shi Y, Boudouh T, Grunder O (2019) A robust optimization for a home health care routing and scheduling problem with consideration of uncertain travel and service times. *Transportation Research Part E: Logistics and Transportation Review* 128:52–95.

Shiri M, Ahmadizar F, Thiruvady D, Farvaresh H (2023) A sustainable and efficient home health care network design model under uncertainty. *Expert Systems with Applications* 211:118185.

Sickinger S, Kolisch R (2009) The performance of a generalized Bailey–Welch rule for outpatient appointment scheduling under inpatient and emergency demand. *Health Care Management Science* 12:408–419.

Singh S, Gray A, Shepperd S, Stott DJ, Ellis G, Hemsley A, Khanna P, Ramsay S, Schiff R, Tsiachristas A, et al. (2022) Is comprehensive geriatric assessment hospital at home a cost-effective alternative to hospital admission for older people? *Age and Ageing* 51(1):afab220.

Srinivas S, Ravindran AR (2018) Optimizing outpatient appointment system using machine learning algorithms and scheduling rules: A prescriptive analytics framework. *Expert Systems with Applications* 102:245–261.

Statistics Canada (2023) Table 14-10-0308-01 employees working overtime. Accessed: March 31, 2024. https://doi.org/10.25318/1410030801-eng.

Tanoumand N, Ünlüyurt T (2021) An exact algorithm for the resource constrained home health care vehicle routing problem. *Annals of Operations Research* 304:397–425.

Taylor SA, Baker TL (1994) An assessment of the relationship between service quality and customer satisfaction in the formation of consumers' purchase intentions. *Journal of Retailing* 70(2):163–178.

Tilk C, Drexl M, Irnich S (2019) Nested branch-and-price-and-cut for vehicle routing problems with multiple resource interdependencies. *European Journal of Operational Research* 276(2):549–565.

Trautsamwieser A, Hirsch P (2014) A branch-price-and-cut approach for solving the medium-term home health care planning problem. *Networks* 64(3):143–159.

Vanderbeck F (2001) A nested decomposition approach to a three-stage, two-dimensional cutting-stock problem. *Management Science* 47(6):864–879.

Welch J, Bailey NJ (1952) Appointment systems in hospital outpatient departments. *The Lancet* 259(6718):1105–1108.

Yin Y, Liu X, Chu F, Wang D (2023) An exact algorithm for the home health care routing and scheduling with electric vehicles and synergistic-transport mode. *Annals of Operations Research* 1–36.

Yuan B, Liu R, Jiang Z (2014) Home health care crew scheduling and routing problem with stochastic service times. *2014 IEEE International Conference on Automation Science and Engineering*, 564–569 (IEEE).

Yuan B, Liu R, Jiang Z (2015) A branch-and-price algorithm for the home health care scheduling and routing problem with stochastic service times and skill requirements. *International Journal of Production Research* 53(24):7450–7464.

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

Zacharias C, Yunes T (2020) Multimodularity in the stochastic appointment scheduling problem with discrete arrival epochs. *Management Science* 66(2):744–763.

Zhang Y, Shen S, Erdogan SA (2017) Distributionally robust appointment scheduling with moment-based ambiguity set. *Operations Research Letters* 45(2):139–144.

Zheng B, Yoon SW, Khasawneh MT, et al. (2015) An overbooking scheduling model for outpatient appointments in a multi-provider clinic. *Operations Research for Health Care* 6:1–10.

Zheng C, Wang S, Li N, Wu Y (2021) Stochastic joint homecare service and capacity

planning with nested decomposition approaches. *European Journal of Operational Research* 295(1):203–222.

Zhou M, Loke GG, Bandi C, Liau ZQG, Wang W (2022) Intraday scheduling with patient re-entries and variability in behaviours. *Manufacturing & Service Operations Management* 24(1):561–579.

A. Appendix

A.1. Robust Appointment Scheduling with Waiting Time Guarantees

A.1.1. Proofs.

Theorem 1 ($\mathcal{NP}$-Hardness, Page 17)**.** *Let all patients have identical waiting time guarantees W. Moreover, let the idle time costs fulfill $c_1 = \ldots = c_\sigma > c_{\sigma+1} = \ldots = c_n$ for some $\sigma \in [n-1]$ and the overtime cost $c_o = 0$. Then the Robust Appointment Scheduling Problem with Waiting Time Guarantees is $\mathcal{NP}$-hard.*

Proof. To simplify the notation, let ϵ be a real number such that $L = \sum_{j=1}^{n} \bar{p}_j - W + \epsilon =: A_{n+1}$.

First, we employ Theorem 2 to reformulate the objective function as

$$\min_{\pi, A} \left\{ \max_{i^* \in \{0,\ldots,n+1\}} \sum_{i=1}^{i^*} c_i (A_i - \underline{C}_{i-1}(\pi, A))^+ + c_o \left(A_{i^*} + \sum_{k=l^*}^{n} \sum_{j=1}^{n} \pi_{kj} \bar{p}_j - L \right)^+ \right\}.$$

Since the overtime cost fulfill $c_o = 0$, it suffices to consider $i^* = n+1$. Moreover, using Theorem 4, the appointment start times can be calculated as $A_1 = 0$, $A_i = (\sum_{l=1}^{i-1} \sum_{j=1}^{n} \pi_{lj} \bar{p}_j - W)^+$ for $i \in \{2, \ldots, n\}$. Therefore, the worst-case cost

equals:

$$
\begin{aligned}
&\sum_{i=2}^{n+1} c_i \left(A_i - \underline{C}_{i-1}(\pi, A)\right)^+ \\
=&\sum_{i=2}^{n+1} c_i \left(\begin{array}{l} \displaystyle\sum_{k=1}^{i-1}\sum_{j=1}^{n} \pi_{kj}\bar{p}_j - W \\ \displaystyle - \max\left\{ \sum_{k=1}^{i-1}\sum_{j=1}^{n} \pi_{kj}\underline{p}_j, \max_{l\in[i-1]} \left(\sum_{k=1}^{l-1}\sum_{j=1}^{n} \pi_{kj}\bar{p}_j + \sum_{k=l}^{i-1}\sum_{j=1}^{n} \pi_{kj}\underline{p}_j \right) - W \right\} \end{array} \right)^+ \\
&+ c_{n+1}\epsilon \\
=&\sum_{i=2}^{n+1} c_i \underbrace{\left(\min\left\{ \sum_{k=1}^{i-1}\sum_{j=1}^{n} \pi_{kj}(\bar{p}_j - \underline{p}_j) - W, \sum_{j=1}^{n} \pi_{i-1,j}(\bar{p}_j - \underline{p}_j) \right\} \right)^+}_{:=s_i} + c_{n+1}\epsilon
\end{aligned}
$$

We observe that $\sum_{j=1}^{n} \pi_{i-1,j}(\bar{p}_j - \underline{p}_j) \geq 0$ holds for all $i \in [2, n+1]$ in the second term in s_i and that $s_i = 0$ if $\sum_{k=1}^{i-1}\sum_{j=1}^{n} \pi_{kj}(\bar{p}_j - \underline{p}_j) - W \leq 0$. Let l^* be the smallest index such that $\sum_{k=1}^{l^*-1}\sum_{j=1}^{n} \pi_{kj}(\bar{p}_j - \underline{p}_j) - W \geq 0$. Then, for all $i > l^*$, it holds that $\sum_{j=1}^{n} \pi_{i-1,j}(\bar{p}_j - \underline{p}_j) \leq \sum_{k=1}^{l^*-1}\sum_{j=1}^{n} \pi_{kj}(\bar{p}_j - \underline{p}_j) - W + \sum_{j=1}^{n} \pi_{i-1,j}(\bar{p}_j - \underline{p}_j) \leq \sum_{k=1}^{i-1}\sum_{j=1}^{n} \pi_{kj}(\bar{p}_j - \underline{p}_j) - W$. Therefore, the worst-case cost finally becomes:

$$
\begin{aligned}
&\sum_{i=1}^{n+1} c_i (A_i - \underline{C}_{i-1})^+ && \text{(A.1)} \\
=&c_{l^*} \left(\sum_{i=1}^{l^*-1}\sum_{j=1}^{n} \pi_{ij}(\bar{p}_j - \underline{p}_j) - W \right) + \sum_{i=l^*+1}^{n+1} c_i \sum_{j=1}^{n} \pi_{i-1,j}(\bar{p}_j - \underline{p}_j) + c_{n+1}\epsilon. && \text{(A.2)}
\end{aligned}
$$

Now, we observe that the index l^* is uniquely determined by the two conditions:

(i) $\sum_{i=1}^{l^*-1}\sum_{j=1}^{n} \pi_{ij}(\bar{p}_j - \underline{p}_j) - W \geq 0$

(ii) $\sum_{i=1}^{l^*-2}\sum_{j=1}^{n} \pi_{ij}(\bar{p}_j - p_j) - W < 0$

Notably, even if $\sum_{i=1}^{l^*-2}\sum_{j=1}^{n} \pi_{ij}(\bar{p}_j - \underline{p}_j) - W = 0$, then the worst-case sum of idle time equals (60).

Therefore, worst-case idle times arise in two ways: If no idle time occurs up to and including appointment $l^* - 1$, then the idle time before appointment l^* equals $\sum_{i=1}^{l^*-1} \sum_{j=1}^{n} \pi_{ij}(\bar{p}_j - \underline{p}_j) - W$. The idle time before all subsequent appointments $i > l^*$ equals $\sum_{j=1}^{n} \pi_{i-1,j}(\bar{p}_j - \underline{p}_j)$. In the case of $l^* = \sigma$, the unit idle time cost before appointment l^* exceeds all later unit costs of idle time. The challenge lies in partitioning the patients into two groups such that the idle time sum $\sum_{i=1}^{l^*-1} \sum_{j=1}^{n} \pi_{ij}(\bar{p}_j - \underline{p}_j) - W$ is minimized. The RASWTG problem under consideration becomes:

$$\min_{l^* \in [2,n+1]} \min_{\pi} \quad c_{l^*} \left(\sum_{j=1}^{n} \sum_{i=1}^{l^*-1} \pi_{ij}(\bar{p}_j - \underline{p}_j) \right) + \sum_{j=1}^{n} \sum_{i=l^*+1}^{n} \pi_{i-1,j} c_i (\bar{p}_j - \underline{p}_j) \tag{A.3}$$

$$- c_{l^*} W + c_{n+1} \epsilon \tag{A.4}$$

$$\text{s.t.} \quad W \leq \sum_{i=1}^{l^*-1} \sum_{j=1}^{n} \pi_{ij}(\bar{p}_j - \underline{p}_j) \leq W + \sum_{j=1}^{n} \pi_{l^*-1,j}(\bar{p}_j - \underline{p}_j) \tag{A.5}$$

$$\sum_{j=1}^{n} \pi_{ij} = 1, \sum_{i=1}^{n} \pi_{ij} = 1 \tag{A.6}$$

$$\pi \in \{0,1\}^{n \times n} \tag{A.7}$$

We now reduce the following $\mathcal{NP}$-hard Subset Sum Problem (Garey and Johnson 1979, p. 226) to a variant of (A.4) – (A.7).

$$\max_{z} \quad \sum_{j=1}^{n} \alpha_j z_j \tag{A.8}$$

$$\text{s.t.} \quad \sum_{j=1}^{n} \alpha_j z_j \leq \beta \tag{A.9}$$

$$z \in \{0,1\}^n \tag{A.10}$$

Defining $y_j := 1 - z_j$ for all $j \in [n]$, the Subset Sum Problem is equivalent to the following problem. In particular, denoting the optimal value of the Subset Sum Problem by OPT_z and the optimal value of the following problem by OPT_y,

it holds $OPT_z = \sum_{j=1}^{n} \alpha_j - OPT_y$.

$$\min_{l^* \in [0,n]} \min_{y} \quad \sum_{j=1}^{n} \alpha_j y_j \tag{A.11}$$

$$\text{s.t.} \quad \sum_{j=1}^{n} \alpha_j - \beta \leq \sum_{j=1}^{n} \alpha_j y_j \tag{A.12}$$

$$\sum_{j=1}^{n} y_j = l^* \tag{A.13}$$

$$y \in \{0,1\}^n \tag{A.14}$$

If the l^*th problem is infeasible, we assign it a huge number. Now, w.l.o.g. (the case $l^* = 0$ is trivial to check), the latter problem is equivalent to:

$$\min_{l^* \in [n]} \min_{\pi} \quad \sum_{i=1}^{l^*} \sum_{j=1}^{n} \alpha_j \pi_{ij} \tag{A.15}$$

$$\text{s.t.} \quad \sum_{j=1}^{n} \alpha_j - \beta \leq \sum_{i=1}^{l^*} \sum_{j=1}^{n} \alpha_j \pi_{ij} \tag{A.16}$$

$$\sum_{j=1}^{n} \pi_{ij} = 1, \sum_{i=1}^{n} \pi_{ij} = 1 \tag{A.17}$$

$$\pi_{ij} \in \{0,1\}^{n \times n} \tag{A.18}$$

We can now use the similarity of (A.15) – (A.18) and (A.4) – (A.7). Let α and β be an instance for (A.15) – (A.18). For $l^* \in [n]$ choose $\bar{p}$, $\underline{p}$ fulfilling $\bar{p}_j - \underline{p}_j := \alpha_j$ for each $j \in [n]$, $W := \sum_{j=1}^{n} \alpha_j - \beta$, $c_1 = \ldots = c_{l^*} := 1$, and $c_{l^*+1} = \ldots = c_{n+1} := 0$. Then solving the l^*th subproblem of (A.15) – (A.18) is equivalent to solving the respective (l^*+1)-st subproblem of (A.4) – (A.5). In particular, an optimal solution (i_{OPT}, π_{OPT}) also fulfills the second inequality in (A.5): If for some $j \in [n]$ and $l \leq i_{OPT}$ such that $\pi_{lj} = 1$ holds $\sum_{i \in [1, i_{OPT}] \setminus \{l\}} \sum_{j=1}^{n} \alpha_j \pi_{ij} > \sum_{j=1}^{n} \alpha_j - \beta$, then there exists a solution (l^*, π^*), $l^* < i_{OPT}$ with smaller solution value, which contradicts the optimality of (i_{OPT}, π_{OPT}).

Thus, Problem (A.4) – (A.7), and therefore also the original problem, are $\mathcal{NP}$-hard. □

A.1.2. Computational Study Details

A.1.2.1. Instances

Figure A.1 presents a boxplot illustrating the service times for the most common examination types.

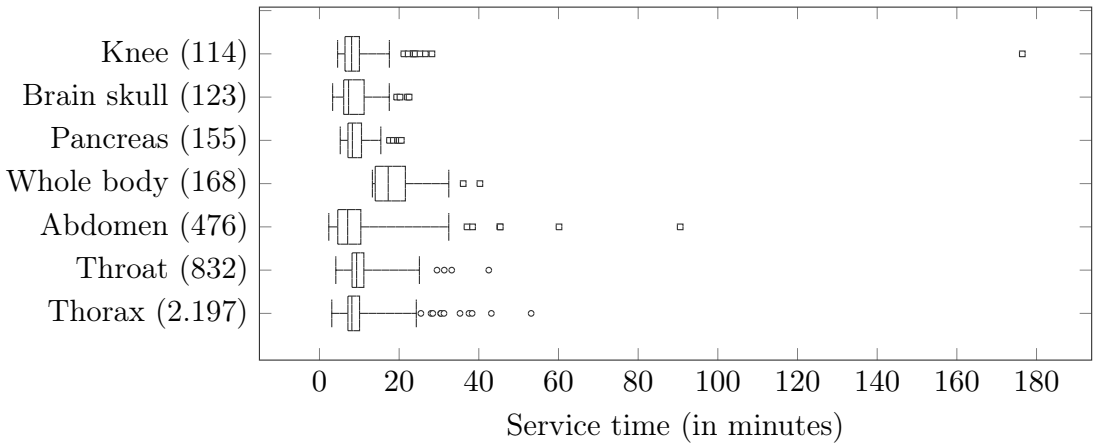

Figure A.1. Service times by examination type
(number of observations in parentheses)

A.1.2.2. Utilitarian Approach

We briefly explore optimal interappointment times to enhance our understanding of optimal WSRAS schedules, as illustrated in Figure A.2. We again compute the interappointment times for ten identical customers with a minimum service time of 15 minutes, a maximum service time of 25 minutes, and a waiting time guarantee of 30 minutes. In contrast to RASWTG, WSRAS interappointment times start at a higher level. Then, depending on the waiting time weight, they gradually decrease until a certain point. For small waiting time weights, most interappointment times match the minimum service time of 15 minutes, leading to little idle time but exploding waiting times. Conversely, higher waiting time weights yield interappointment times closer to the maximum service time of 25 minutes, resulting in short waiting times but extended idle time. Intriguingly, when $c_w = 0.1$, WSRAS accumulates waiting times until, at a certain point, they are alleviated by a single long interappointment time.

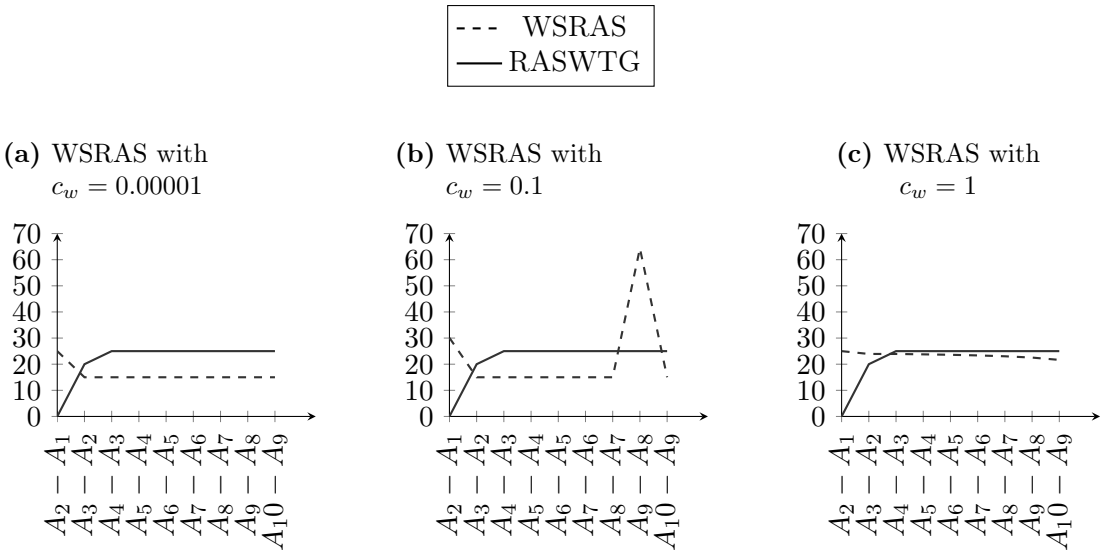

Figure A.2. Optimal interappointment times (in minutes)

Finally, we note that the WSRAS problem with constant idle time costs closely resembles problem (1) in Schulz and Udwani (2019), except that the latter does not consider a fixed planning horizon L. While initially minimizing a weighted sum of idle times and waiting time costs under a collective worst-case scenario, Schulz and Udwani (2019) demonstrate the sufficiency of considering the same $n + 1$ scenarios as in WSRAS.

Moreover, comparing an RASWTG-optimal schedule that does not consider a fixed planning horizon (e.g., $c_{n+1} = c_o = 0$) to a schedule computed with the optimal scheduling rule in Schulz and Udwani (2019) and under the respective assumptions, we find that RASWTG improves capacity utilization for sufficiently long waiting time guarantees. In particular, it is easy to show that the RASWTG schedule reduces worst-case idle time costs in the case of $W > \sum_{i=1}^{n-1} \sum_{j=1}^{n} \pi_{ij} \frac{u}{u+o_{\geq i}} \left(\bar{p}_j - \underline{p}_j\right)$.

A.2. Robust Predictive Overbooking

A.2.1. Proofs

Lemma 3. For any service time scenario $p \in \mathcal{P}$ and no-show scenario $\lambda \in \Lambda$, the completion time of appointment $i \in [n]$ equals

$$C_i = \max_{l \in [1,i]} \left(A_l + \sum_{k=l}^{i} \sum_{j=1}^{n} \pi_{kj} \lambda_j p_j \right). \tag{A.19}$$

Proof. Proof of Lemma 3. We prove by induction that the recursive computation of the completion times $\max\{A_i, C_{i-1}\} + \sum_{j=1}^{n} \pi_{ij} \lambda_j p_j$ is equivalent to the maximization problem $\max_{l \in [1,i]} \left(A_l + \sum_{\sigma=l}^{i} \sum_{j=1}^{n} \pi_{\sigma j} \lambda_j p_j \right)$. For $i = 1$, the terms are equivalent. Suppose the claim is true for all appointments $1, \ldots, i-1$. Then the recursive definition of the completion time for appointment i is equivalent to $\max \left\{ A_i, \max_{l \in [1,i-1]} \left(A_l + \sum_{\sigma=l}^{i-1} \sum_{j=1}^{n} \pi_{\sigma j} \lambda_j p_j \right) \right\} + \sum_{j=1}^{n} \pi_{ij} \lambda_j p_j$. Taking the last summand into the maximization problem renders this equivalent to the above maximization problem. □

A.2.2. Linearization

For any given but fixed $i^* \in \{0, \ldots, n+1\}$, we compute each appointment i's cost function $c(i)$ from Theorem 5 via $c(i) = \sum_{l=1}^{n+1} \gamma_{il} c_l$ with $c_{n+1} := \max_{l \in [n]} c_l$ using any feasible solution $\gamma \in \{0,1\}^{n \times (n+1)}$ of the following mixed-integer linear problem.

$$\gamma_{il} \leq \Delta_l M \qquad \forall l \in [n] \forall i \in [n+1] \tag{A.20}$$

$$\sum_{z_{li}=i}^{l-1} \Delta_{z_{li}} \leq L(1 - \gamma_{il}) \qquad \forall z_{li} \in [n] \forall i \in [n+1] \tag{A.21}$$

$$\sum_{l=1}^{i-1} \gamma_{il} = 0 \qquad \forall i \in [n] \tag{A.22}$$

$$\sum_{l=i}^{n+1} \gamma_{il} = 1 \qquad \forall i \in [n] \tag{A.23}$$

Constraints (A.20) ensure that, if there is no positive amount of idle time before an appointment $l \in [n]$, then $\gamma_{il} = 0$. We carefully choose $M \geq \frac{1}{\min_{l \in [n]}\{\Delta_l : \Delta_l > 0\}}$ to ensure numerical stability. An alternative approach involves replacing (A.20) with an indicator constraint, namely $\Delta_l = 0 \Rightarrow \gamma_{il} = 0$. Constraints (A.22) and (A.23) ensure that each appointment i is assigned a unit idle time cost c_l with $l \geq i$. Constraints (A.21) ensure that, if $\gamma_z = 1$, then there occurs no idle time before any of the appointments $i, ..., z_{li} - 1$. Therefore, Constraints (A.20) to (A.23) ensure that $\gamma_{il} = 1$ holds if and only if $l \in [i, n]$ is the next appointment before which a positive amount of idle time occurs, and $\gamma_{i,n+1} = 1$ if there is no such appointment.

A.3. Robust Routing and Scheduling of Home Healthcare Workers

A.3.1. Dualization Scheme for Adversarial Problems

To apply the classical dualization scheme as proposed by Bertsimas and Sim (2004), we reformulate the service start times. In particular, we replace the binary variable x with a new binary variable $z_{j'j}^{k,d,i} \in \{0,1\}$, which additionally indicates the sequence of patients treated by caregiver k on day d. Specifically, $z_{j'j}^{kdi}$ takes a value of 1 if and only if the ith patient treated by caregiver k on day d is patient j, and the patient treated before j was patient j'. Notably, we keep both patient indices due to the consideration of travel times. Now, we shift the focus from the start time of patient i to the start time of the ith patient treated by caregiver k on day d.

Lemma 4 (Start Time Reformulation)**.** Let $\bar{p}_0 := 0$ and $\hat{p}_0 := 0$. For a fixed caregiver and day, consider the binary variable $z_{j',j}^i$, where $z_{j',j}^i$ equals 1 if the ith visit is with patient $j \in \mathcal{I}$ and patient $j' \in \mathcal{I} \cup \{0\}$ was visited immediately before, and 0 otherwise. Then, the service start times can be computed as follows.

a. Compute the service start time s_i for patient $i \in \mathcal{I}$, initializing s_0 as $\underline{W}$, as follows:

$$s_i \geq \max\left\{\underline{T}_i, \max_{j \in \mathcal{I} \cup \{0\}} \left(s_j + \bar{p}_i + \gamma_i^p \hat{p}_i + \bar{t}_{ij} + \gamma_{ij}^t \hat{t}_{ij}\right) x_{ji}\right\} \tag{A.24}$$

b. Compute the service start time $\tilde{s}_i$ for the ith visited patient, with $\underline{T}_0 := \underline{W}$, and $\phi(\gamma) := \bar{p}_j + \gamma_j^p \hat{p}_j + \bar{t}_{jj'} + \gamma_{jj'}^t \hat{t}_{jj'}$, as follows:

$$\tilde{s}_i \geq \max\left\{\sum_{j \in \mathcal{I}} \sum_{\substack{j' \in \mathcal{I} \\ \cup\{0\}}} z_{j'j}^i \underline{T}_j, \max_{i' \leq i}\left\{\sum_{j \in \mathcal{I}} \sum_{\substack{j' \in \mathcal{I} \\ \cup\{0\}}} z_{j'j}^{i'} \underline{T}_{j'} + \sum_{l=i'}^{i} \sum_{j \in \mathcal{I}} \sum_{\substack{j' \in \mathcal{I} \\ \cup\{0\}}} \phi(\gamma) z_{j'j}^l\right\}\right\} \tag{A.25}$$

Proof. Proof of Lemma 4. Let $\tilde{s}_0 := \underline{W}$. We prove by induction that (A.25) is equivalent to the recursive formulation of start times

$$\tilde{s}_i \geq \max \left\{ \sum_{j \in \mathcal{I}} \sum_{j' \in \mathcal{I} \cup \{0\}} z^i_{j'j} \underline{T}_j, \tilde{s}_{i-1} + \sum_{j \in \mathcal{I}} \sum_{j' \in \mathcal{I} \cup \{0\}} \phi(\gamma) z^i_{j'j} \right\}. \tag{A.26}$$

For $i = 1$, the terms are equivalent. Suppose the claim is true for all visits $1, \ldots, i-1$. Then replacing $\tilde{s}_{i-1}$ in (A.26) with the formulation given by (A.25) shows the equivalence. It is evident that computing the service start times using (A.26) is equivalent to using (A.24). □

Let n_{max} denote the maximum number of patients treated by any caregiver on any day. Using the binary variable z and defining $p_0 := 0$, $\hat{p}_0 := 0$, $\bar{p}_0 := 0$, we can reformulate Constraints (4.2) as:

$$s^{kd}_{i^*} \leq \sum_{j=1}^{n} \sum_{j'=1}^{n} \bar{T}_{jd} z^{k,d,i^*}_{j'j}$$
$$\forall k \in \mathcal{K} \, \forall d \in \mathcal{D} \, \forall i^* \in [n_{max}] \tag{A.27}$$
$$s^{kd}_{i^*} \geq \underline{W}_{kd} + \max_{\substack{p \in \mathcal{U}^p \\ t \in \mathcal{U}^t}} \left\{ \sum_{i=1}^{i^*} \sum_{j=1}^{n} \sum_{j'=0}^{n} (p_j + t_{j'j}) z^{k,d,i}_{j'j} \right\}$$
$$\forall k \in \mathcal{K} \, \forall d \in \mathcal{D} \, \forall i^* \in [n_{max}] \tag{A.28}$$
$$s^{kd}_{i^*} \geq \sum_{j=1}^{n} \sum_{j'=0}^{n} \underline{T}_{jd} z^{k,d,i'}_{j'j} + \max_{\substack{p \in \mathcal{U}^p \\ t \in \mathcal{U}^t}} \left\{ \sum_{i=i'}^{i^*} \sum_{j=1}^{n} \sum_{j'=0}^{n} (p_j + t_{j'j}) z^{k,d,i}_{j'j} \right\}$$
$$\forall k \in \mathcal{K} \, \forall d \in \mathcal{D} \, \forall i^* \in [n_{max}] \forall i' \in [i^*] \tag{A.29}$$

Now, replacing the maximization problems in A.28 and A.29 by their LP duals, we can replace Constraints (4.2) by the following linear formulation:

$$s_{i^*}^{kd} \leq \sum_{j=1}^{n}\sum_{j'=1}^{n} \bar{T}_{jd} z_{j'j}^{k,d,i^*}$$

$$\forall k \in \mathcal{K} \forall d \in \mathcal{D} \forall i^* \in [n_{max}] \tag{A.30}$$

$$s_{i^*}^{kd} \geq \underline{W}_{kd} + \sum_{i=1}^{i^*}\sum_{j=1}^{n}\sum_{j'=0}^{n} (\bar{p}_j + \bar{t}_{j'j}) z_{j'j}^{k,d,i} + \Gamma^p y_p^{(k,d,i^*,0)} + \Gamma^t y_t^{(k,d,i^*,0)} + \sum_{j=1}^{n} (y^p)_j^{(k,d,i^*,0)} + \sum_{j=1}^{n}\sum_{j'=0}^{n} (y^t)_{j'j}^{(k,d,i^*,0)}$$

$$\forall k \in \mathcal{K} \forall d \in \mathcal{D} \forall i^* \in [n_{max}] \tag{A.31}$$

$$y_p^{(k,d,i^*,0)} + (y^p)_j^{(k,d,i^*,0)} \geq \sum_{i=1}^{i^*}\sum_{j'=1}^{n} \hat{p}_j z_{j'j}^{k,d,i}$$

$$\forall k \in \mathcal{K} \forall d \in \mathcal{D} \forall i^* \in [n_{max}] \forall j \in N \tag{A.32}$$

$$y_t^{(k,d,i^*,0)} + (y^p)_{j'j}^{(k,d,i^*,0)} \geq \sum_{i=1}^{i^*} \hat{t}_{j'j} z_{j'j}^{k,d,i}$$

$$\forall k \in \mathcal{K} \forall d \in \mathcal{D} \forall i^* \in [n_{max}] \forall j' \in N_0 \forall j \in N \tag{A.33}$$

$$s_{i^*}^{kd} \geq \sum_{j=1}^{n}\sum_{j'=0}^{n} \underline{T}_{jd} z_{j'j}^{k,d,i'} + \sum_{i=i'}^{i^*}\sum_{j=1}^{n}\sum_{j'=0}^{n} (\bar{p}_j + \bar{t}_{j'j}) z_{j'j}^{k,d,i} + \Gamma^p y_p^{(k,d,i^*,i')} + \Gamma^t y_t^{(k,d,i^*,i')} + \sum_{j=1}^{n} (y^p)_j^{(k,d,i^*,i')} + \sum_{j=1}^{n}\sum_{j'=0}^{n} (y^t)_{j'j}^{(k,d,i^*,i')}$$

$$\forall k \in \mathcal{K} \forall d \in \mathcal{D} \forall i^* \in [n_{max}] \forall i' \in [i^*] \tag{A.34}$$

$$y_p^{(k,d,i^*,i')} + (y^p)_j^{(k,d,i^*,i')} \geq \sum_{i=i'}^{i^*}\sum_{j'=1}^{n} \hat{p}_j z_{j'j}^{k,d,i}$$

$$\forall k \in \mathcal{K} \forall d \in \mathcal{D} \forall i^* \in [n_{max}] \forall i' \in [i^*] \forall j \in N \tag{A.35}$$

$$y_t^{(k,d,i^*,i')} + (y^p)_{j'j}^{(k,d,i^*,i')} \geq \sum_{i=i'}^{i^*} \hat{t}_{j'j} z_{j'j}^{k,d,i}$$

$$\forall k \in \mathcal{K} \forall d \in \mathcal{D} \forall i^* \in [n_{max}] \tag{A.36}$$

$$\forall i' \in [i^*] \forall j' \in N_0 \forall j \in N \tag{A.37}$$

$$y \geq 0 \tag{A.38}$$

This formulation requires $|\mathcal{I}|^2 \cdot |\mathcal{K}| \cdot |\mathcal{D}| \cdot n_{max}$ binary variables, $\mathcal{O}\left(|\mathcal{I}|^2 \cdot |\mathcal{K}| \cdot |\mathcal{D}| \cdot (n_{max})^2\right)$ new continuous variables and replaces Constraints (4.2) with $\mathcal{O}\left(|\mathcal{I}|^2 \cdot |\mathcal{K}| \cdot |\mathcal{D}| \cdot (n_{max})^2\right)$ new constraints.

A.3.2. Nested Branch-and-Price Procedure

In the following, we present the LP duals for the linear relaxation of both the first- and second-level master problems, along with the corresponding pricing problems.

A.3.2.1. LP Duals

The LP dual of the linear relaxation of the first-level master problem (4.20) – (4.23) can be stated as follows:

$$\min \sum_{k \in \mathcal{K}} v_k + \sum_{i \in \mathcal{I}_R} w_i \tag{A.39}$$

$$\text{s.t. } v_k + \sum_{i \in \mathcal{I}_R} a^i_{k,P}\, w_i \geq Rev_{k,P} - C(T)_{k,P} - C(W)_{k,P} \quad \forall k \in \mathcal{K} \forall P \in \mathcal{P}_k \tag{A.40}$$

$$w \geq 0 \tag{A.41}$$

The dual perspective provides valuable insights into the problem. It aims to find an optimal value for each caregiver and each new patient. For each weekly plan, the total value generated by the caregiver and the newly accepted patients he or she visits must be at least as high as the profit generated by the weekly plan. The LP dual aims to minimize the aggregate value associated with all caregivers and new patients.

The LP dual of the linear relaxation of the second-level master problem (4.25) – (4.30) can be stated as follows:

$$\min \sum_{d\in\mathcal{D}} u_d + \sum_{i\in\mathcal{I}^{\mathcal{R}}} v_i z_i + \sum_{i\in\mathcal{I}^{\mathcal{R}}} \sum_{\tilde{d}\in[|\mathcal{D}|-d_i]_0\cap[d_i-d]_0} q_{i\tilde{d}} \tag{A.42}$$

$$\text{s.t. } u_d + \sum_{i\in\mathcal{I}^{\mathcal{R}}} a^i_{d,R} z_i + \sum_{i\in\mathcal{I}^{\mathcal{R}}} \sum_{\tilde{d}\in[|\mathcal{D}|-d_i]_0\cap[d_i-d]_0} a^i_{d,R} q_{i\tilde{d}} \tag{A.43}$$

$$- \sum_{i\in\mathcal{I}^{\mathcal{R}}} \sum_{\tilde{d}\in\mathcal{D}} a^i_{d,R} y_{i\tilde{d}} + \sum_{i\in\mathcal{I}^{\mathcal{R}}} v_i a^i_{d,R} y_{id} \tag{A.44}$$

$$\geq Rev_{d,R} - C(T)_{d,R} - C(W)_{d,R} - \sum_{i\in\mathcal{I}^{\mathcal{R}}} \frac{w^*_i}{v_i} a^i_{d,R} \quad \forall d\in\mathcal{D}\,\forall R\in\mathcal{R}_d \tag{A.45}$$

$$z, y, q \geq 0 \tag{A.46}$$

The second-level dual LP evaluates the constraints of a single route per day (u), the fulfillment of the exact number of required visits for each new patient (z, y), and the adherence to the minimum required time period between two consecutive visits (q).

A.3.2.2. Pricing Problems

The first-level pricing problem is solved independently for each fixed caregiver $k \in \mathcal{K}$. It generates a feasible weekly plan that maximizes the profit, minus the value of the newly accepted patients (determined by the dual value w^*). All constraints are known from Constraints ((2)) –((11)) and Contraints (4.15) – (4.19).

$$\max \sum_{i\in\mathcal{I}} R_{ki} v_i \delta_i - \sum_{d\in\mathcal{D}} \sum_{i,j\in\mathcal{I}} C^T_{dij} x^d_{ij} - C^W_k \sum_{d\in\mathcal{D}} \left(s^d_{n+1} - \underline{W}_{kd} \right) - \sum_{i\in\mathcal{I}^{\mathcal{R}}} w^*_i \, \delta_i \tag{A.47}$$

$$\text{s.t. } s^d_{i,\gamma^p,\gamma^t} \in [\underline{T}_{id}, \bar{T}_{id}] \quad \forall d \in \mathcal{D} \, \forall i \in \mathcal{I} \, \forall \gamma^p \in [\Gamma^p]_0 \, \forall \gamma^t \in [\Gamma^t]_0 \tag{A.48}$$

$$s^d_{n+1} \in [\underline{W}_{kd}, \bar{W}_{kd}] \quad \forall d \in \mathcal{D} \tag{A.49}$$

$$s^d_{i,0,\zeta^t} \geq \left(\underline{W}_{kd} + \bar{t}_{0i} + \zeta^t \hat{t}_{0i} \right) x^d_{0i} \quad \forall d \in \mathcal{D} \, \forall i \in \mathcal{I} \, \forall \zeta^t \in \{0,1\} \tag{A.50}$$

$$s^d_{i,\gamma^p+\zeta^p,\gamma^t+\zeta^t} \geq s^d_{j,\gamma^p,\gamma^t} + (\bar{p}_j + \zeta^p \hat{p}_j + \bar{t}_{ji} + \zeta^t \hat{t}_{ji}) x^d_{ji} + (\underline{T}_{id} - \bar{T}_{jd})(1 - x^d_{ji}) \quad \forall d \in \mathcal{D} \, \forall i,j \in \mathcal{I} \, \forall \gamma^p \in [\Gamma^p]_0 \, \forall \gamma^t \in [\Gamma^t]_0 \quad \forall \zeta^p, \zeta^t \in \{0,1\} : \gamma^p + \zeta^p \leq \Gamma^p, \gamma^t + \zeta^t \leq \Gamma^t \tag{A.51}$$

$$s^d_{n+1} \geq s^d_{i,\Gamma^p-\zeta^p,\Gamma^t-\zeta^t} + (\bar{p}_i + \zeta^p \hat{p}_i + \bar{t}_{i,n+1} + \zeta^t \hat{t}_{i,n+1}) x^d_{i,n+1} + (\underline{W}_{kd} - \bar{T}_{id})(1 - x^d_{i,n+1}) \quad \forall d \in \mathcal{D} \, \forall i \in \mathcal{I} \forall \zeta^p, \zeta^t \in \{0,1\} \tag{A.52}$$

$$\sum_{d\in\mathcal{D}} \sum_{j\in\mathcal{I}} x^d_{ij} = v_i \delta_i \qquad \forall i \in \mathcal{I} \tag{A.53}$$

$$\delta_i = 1 \qquad \forall i \in \mathcal{I}^E_k \tag{A.54}$$

$$\sum_{j\in\mathcal{I}} x^d_{ij} \leq C^{kd}_i \qquad \forall d \in \mathcal{D} \, \forall i \in \mathcal{I}^{\mathcal{R}} \tag{A.55}$$

$$\sum_{j\in\mathcal{I}} x^d_{ij} + \sum_{\mu\in[d_i]_0} \sum_{j\in\mathcal{I}} x^{d+\mu}_{ij} \leq 1 \qquad \forall d \in \mathcal{D} \, \forall i \in \mathcal{I}^{\mathcal{R}} \tag{A.56}$$

$$\sum_{j\in\mathcal{J}} x^d_{ji} - \sum_{j\in\mathcal{J}} x^d_{ij} = 0, \sum_{i\in\mathcal{J}} x^d_{ii} = 0 \qquad \forall d \in \mathcal{D} \, \forall i \in \mathcal{I} \tag{A.57}$$

$$\sum_{i\in[n+1]} x^d_{0,i} = 1, \sum_{i\in N_0} x^d_{i,n+1} = 1 \qquad \forall d \in \mathcal{D} \tag{A.58}$$

$$x^d_{ij}, \delta_i \in \{0,1\} \qquad \forall d \in \mathcal{D} \, \forall i,j \in [n+1]_0 \tag{A.59}$$

$$s^d_{i,\gamma^p,\gamma^t}, s^d_{n+1} \geq 0 \qquad \forall d \in \mathcal{D} \, \forall i \in \mathcal{I} \, \forall \gamma^p \in [\Gamma^p]_0 \, \forall \gamma^t \in [\Gamma^t]_0 \tag{A.60}$$

The first-level pricing problem is a mixed-integer linear program with $\mathcal{O}\left(n^2 \cdot |\mathcal{D}|\right)$ binary variables, $\mathcal{O}\left(n \cdot |\mathcal{D}| \cdot |\Gamma^p| \cdot |\Gamma^t|\right)$ continuous variables and $\mathcal{O}(n^2 \cdot |\mathcal{D}| \cdot |\Gamma^p| \cdot |\Gamma^t|)$ constraints.

The second-level pricing problem is solved independently for each fixed caregiver $k \in \mathcal{K}$ and day $d \in \mathcal{D}$. It generates a feasible daily route that maximizes the profit, minus the value of the newly accepted patients (determined by the first- and second-level dual values).

$$\max \sum_{i \in \mathcal{I}_k} R_{ki}\delta_i - \sum_{i,j \in \mathcal{I}_k} C^T_{dij} x_{ij} - C^W_k \left(s_{n+1} - \underline{W}_{kd}\right) \tag{A.61}$$
$$- \left(\sum_{i \in \mathcal{I}^{\mathcal{R}}} \frac{w^*_i}{v_i} + \sum_{i \in \mathcal{I}^{\mathcal{R}}} z^*_i + \sum_{i \in \mathcal{I}^{\mathcal{R}}} \left(\sum_{\tilde{d} \in \mathcal{D}} y^*_{i\tilde{d}} - v_i y^*_{id} \right) + \sum_{i \in \mathcal{I}_k} \sum_{\substack{\mu \in [|\mathcal{D}| - d_i]: \\ \mu \in [d_i - d]_0}} q^*_{i,\mu} \right) \delta_i$$

$$\text{s.t. } s_{i,\gamma^p+\zeta^p,\gamma^t+\zeta^t} \geq s_{j,\gamma^p,\gamma^t} + (\bar{p}_j + \zeta^p \hat{p}_j + \bar{t}_{ji} + \zeta^t \hat{t}_{ji}) x_{ji} + (\underline{T}_{id} - \bar{T}_{jd})(1 - x_{ji}) \quad \forall i, j \in \mathcal{I}_k \, \forall \gamma^p \in [\Gamma^p]_0 \, \forall \gamma^t \in [\Gamma^t]_0 \, \forall \zeta^p, \zeta^t \in \{0,1\} : \gamma^p + \zeta^p \leq \Gamma^p, \gamma^t + \zeta^t \leq \Gamma^t \tag{A.62}$$

$$s_{i,0,\zeta^t} \geq \left(\underline{W}_{kd} + \bar{t}_{0i} + \zeta^t \hat{t}_{0i}\right) x_{0i} + (\underline{W}_{kd} - \bar{T}_{id})(1 - x_{i,n+1}) \quad \forall i \in \mathcal{I}_k \, \forall \zeta^p \zeta^t \in \{0,1\} \tag{A.63}$$

$$s_{i,\gamma^p,\gamma^t} \in [\underline{T}_{id}, \bar{T}_{id}] \quad \forall i \in \mathcal{I}_k \, \forall \gamma^p \in [\Gamma^p]_0 \, \forall \gamma^t \in [\Gamma^t]_0 \tag{A.64}$$

$$s_{n+1} \geq s_{i,\gamma^p-\zeta^p,\Gamma^t-\zeta^t} + (\bar{p}_i + \zeta^p \hat{p}_i + \bar{t}_{i,n+1} + \zeta^t \hat{t}_{i,n+1}) x_{i,n+1} \quad \forall i \in \mathcal{I}_k \, \forall \zeta^t \in \{0,1\} \tag{A.65}$$

$$s_{n+1} \in [\underline{W}_{kd}, \bar{W}_{kd}] \tag{A.66}$$

$$\sum_{j \in \mathcal{I}_k} x_{ij} = \delta_i \qquad \forall i \in \mathcal{I}_k \quad \text{(A.67)}$$

$$\sum_{j \in \mathcal{I}_k} x_{ij} \leq C_i^{kd} \qquad \forall i \in \mathcal{I}_k \quad \text{(A.68)}$$

$$\sum_{j \in \mathcal{I}_k} x_{ji} - \sum_{j \in \mathcal{I}_k} x_{ij} = 0, \sum_{i \in \mathcal{I}_k} x_{ii} = 0 \qquad \forall i \in \mathcal{I}_k \quad \text{(A.69)}$$

$$\sum_{\substack{i \in \mathcal{I}_k \\ \cup \{n+1\}}} x_{0,i} = 1, \sum_{\substack{i \in \mathcal{I}_k \\ \cup \{0\}}} x_{i,n+1} = 1 \quad \text{(A.70)}$$

$$x_{ij}, \delta_i \in \{0,1\} \qquad \forall i,j \in \mathcal{I}_k \cup \{0\} \cup \{n+1\} \quad \text{(A.71)}$$

$$s_{i,\gamma^p,\gamma^t}, s_{n+1} \geq 0 \qquad \forall i \in \mathcal{I}_k \, \forall \gamma^p \in [\Gamma^p]_0 \, \forall \gamma^t \in [\Gamma^t]_0 \quad \text{(A.72)}$$

The second-level pricing problem is a MILP with $\mathcal{O}(n^2)$ binary variables, $\mathcal{O}(n \cdot |\Gamma^p| \cdot |\Gamma^t|)$ continuous variables and $\mathcal{O}(n^2 \cdot |\Gamma^p| \cdot |\Gamma^t|)$ constraints.

A.3.3. Case Study

Appendix A.3.3.1 provides detail on the data and Appendix A.3.3.2 provides additional details on the computational results.

A.3.3.1. Data

The original data set encompasses detailed information on 166,641 visits made by 807 caregivers to 7,405 patients across eight regional branches and five different disciplines. Data was collected by a US national healthcare agency headquartered in Dallas, Texas over a 36-week period from July 21st, 2018 to March 18th, 2019. Figure A.3 illustrates the weekly visit counts for the entire data set. To ensure data completeness, we removed weeks with less than 4,000 visits, resulting in 23 considered weeks. The data comprises information on the patient locations, required services, and timely availability for each patient, as well as the qualification, employment status, and compensation type for each caregiver. Table A.3 summarizes the type of employment and compensation for all caregivers in the data set.

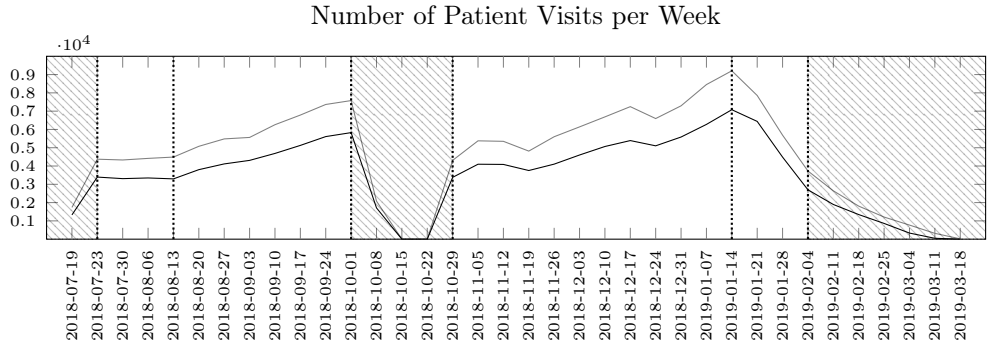

Figure A.3. Temporal trends in weekly patient visits
(overall visits depicted in gray, regular visits in black; shaded regions indicate periods with fewer than 4,000 visits)

Regional branches. The original data set comprises visits from eight regional branches located in four US states. With the exception of seven caregivers, all others exclusively visited patients in a single region. A summary of the eight regions is presented in Table A.1, and a scatter plot of patient locations within each region is shown in Figure A.4. Variability in travel distances and times within each region influences the home healthcare routing and scheduling problem. Accordingly, we categorized the regions into three groups: town-shorttravel, city-midtravel, and town-longtravel. Out of each group, we chose up to one candidate for our instance generation.

Table A.1. Regional branch attributes

	Patients	Travel time (avg.)	Distance (avg.)	Population	Group
Region-T19:	729	19.55 min	21.91 km	$< 500,000$	town-shorttravel
Region-T20:	1,138	20.22 min	22.60 km	$< 100,000$	town-shorttravel
Region-T30:	525	30.82 min	28.98 km	$< 500,000$	town-shorttravel
Region-T33:	1,856	33.94 min	38.03 km	$< 1,000,000$	city-midtravel
Region-T36:	1,677	36.06 min	41.87 km	$< 1,500,000$	city-midtravel
Region-T40:	398	40.64 min	42.63 km	$< 10,000$	town-longtravel
Region-T48:	559	48.06 min	54.18 km	$< 100,000$	town-longtravel
Region-T60:	523	60.98 min	74.69 km	$< 100,000$	town-longtravel

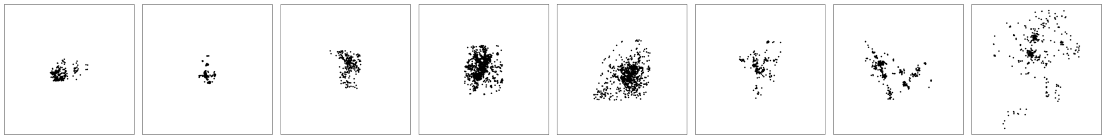

Figure A.4. Scatter plots of patient locations by regional branch
(arranged in increasing order with respect to average travel time)

Services. Patient visits fall into the following disciplines: nursing (45.33% - comprised of skilled nursing [SN], registered nursing [RN], licensed practical nursing [LPN], and psychiatric nursing [PSY]), physical therapy (36.26% - comprised of physical therapy [PT] and physical therapy assistance [PTA]), occupational therapy (12.55% - comprised of occupational therapy [OT] and occupational therapy aid [COTA]), and other categories (5.87% - including speech therapy [ST], home health aid [HHA], and medical social work [MSW]). Qualifications required for the different services vary. For example, PSY services necessitate PSY-qualified nurses, RN services require RN-qualified nurses, while LPN and SN services can be delivered by both LPN- and RN-qualified nurses. The qualification hierarchy is denoted as $RN \succcurlyeq LPN$, $RN \succcurlyeq SN$, $PT \succcurlyeq PTA$, and $OT \succcurlyeq COTA$. The visits are further classified into various types, including routine visits (75.91%), reassessment visits (9.88%), initial visits (7.98%), discharge

Table A.2. Services (excluding 1.78% categorized as 'Other')

	Total visits	Frequency (per patient)		Visit types		Change rate	Wage
		Weekly	LoS	Routine	Other		
SN:	41,889	1 visit(s)	5 weeks	100 %	0 %	0.65	$26.86
RN:	18,361	1 visit(s)	2 weeks	18 %	82%	4.84	$42.80
LPN:	6,959	1 visit(s)	4 weeks	98 %	2 %	0.75	$26.86
PSY:	6,446	1 visit(s)	9 weeks	80 %	20 %	0.18	$59.94
PTA:	34,639	2 visit(s)	5 weeks	100 %	0 %	0.31	$31.01
PT:	25,780	1 visit(s)	3 weeks	46 %	54 %	1.36	$47.10
OT:	12,037	1 visit(s)	2 weeks	49 %	51 %	1.00	$44.61
COTA:	8,869	1.5 visit(s)	3 weeks	100 %	0 %	0.43	$20.71
ST:	6,195	1 visit(s)	4 weeks	71 %	29 %	0.28	$38.01
HHA:	2,502	1 visit(s)	3 weeks	100 %	0 %	0.36	$14.15
MSW:	1,074	1 visit(s)	1 weeks	15 %	85 %	9.07	$30.17

Table A.3. Caregivers

Employment				**Compensation**			
full time	PRN	part time	other	per visit	hourly	salaried	other
51.93%	18.98%	2.49%	26.60%	37.41%	23.57%	12.68%	26.34%

visits (2.62%), and other types of visits (3.62%). In our computational study, we focus exclusively on scheduling routine visits. Tables A.2 summarizes key characteristics of each service that impact the home healthcare routing and scheduling problem, such as mean weekly visits and length of stay (LoS) per patient, visit type distribution, change rate (average ratio of new patients to existing patients), wage (according to of Labor Statistics (2022)), and expected service time.

A.3.3.2. Results

In addition to the results presented in the main part, Table A.4 summarizes the runtimes for all instances, averaged over three weeks. Similarly, Table A.5 summarizes the profit across all instances.

Table A.4. Average runtimes by discipline, region, level of robustness, and time window size (in seconds)

		Physical Therapy			Nursing		
		Wide	Narrow	Tight	Wide	Narrow	Tight
Region T-20	None	10.50	3.83	3.99	39.51	80.57	36.69
	Low	177.67	12.67	15.20	411.04	13.86	42.43
	Medium	1,116.91	205.52	37.06	6,758.96	11.60	434.74
	High	25.67	3.45	3.43	51.86	187.55	11.61
Region T-60	None	9.92	1.33	2.91	21.00	1.39	1.52
	Low	81.65	5.45	9.54	159.62	10.16	11.41
	Medium	260.32	14.03	17.59	373.75	30.28	30.34
	High	4.35	1.27	2.25	4.67	1.24	0.86

Table A.5. Average profit by discipline, region, level of robustness, and time window size (in USD)

		Nursing			Physical Therapy		
		Wide	Narrow	Tight	Wide	Narrow	Tight
Region T-20	None	17,634	17,872	17,625	6,939	6,841	6,841
	Low	17,634	17,625	17,620	6,914	6,817	6,641
	Medium	17,634	17,625	17,620	6,819	6,731	6,624
	High	17,634	17,624	17,620	5,448	5,394	5,391
Region T-60	None	4,525	4,200	4,290	6,202	5,458	5,447
	Low	4,015	3,803	3,325	5,859	5,054	4,944
	Medium	3,723	3,329	3,138	5,361	4,946	4,944
	High	3,458	3,231	3,138	5,041	4,946	4,944

B. Notations

B.1. Abbreviations

Chapters 2 and 3

ASAP	As soon as possible
MILP	Mixed-integer linear program
RASWTG	Robust appointment scheduling with waiting time guarantees
WSRAS	Weighted sum robust appointment scheduling
RPOWTG	Robust predictive overbooking with waiting time guarantees
SVF	Smallest variance first
SVF-WTG	Smallest variance first with waiting time guarantees
WSRAS	Weighted sum robust appointment scheduling

Chapter 4

B&P	Branch and price
HHCR&S	Home healthcare routing and scheduling
LBBD	Logic-based Benders decomposition
MILP	Mixed-integer linear program
R-HHCR&S	Robust home healthcare routing and scheduling
R-TSPTW	Robust traveling salesman problem with time windows

B.2. Symbols

Chapters 2 and 3

$\mathcal{C}$	candidate set of predicted no-shows
c_i	unit cost for idle time before appointment i
c_o	unit cost for overtime
c_w	unit cost for waiting time
L	length of the planning horizon
M	sufficiently large parameter
n	number of patients
$\mathcal{O}$	Bachmann–Landau notation
$\mathcal{P}$	set of service time scenarios
W_j	waiting time guarantee for patient j
Γ^{FN}	uncertainty budget for false negative predictions
Γ^{FP}	uncertainty budget for false positive predictions
Λ	set of no-show scenarios

Chapter 4

C^W	wage cost
C^T	travel cost
$\mathcal{D}$	set of days
D	day pattern
$\mathcal{I}$	set of patients
$\mathcal{I}^R$	set of new patient requests
$\mathcal{I}^E$	set of existing patients
$\mathcal{K}$	set of caregivers
$\mathcal{O}$	Bachmann–Landau notation
$\mathcal{P}_k$	set of all weekly plans for caregiver k
$\tilde{p}$	expected service times
$\hat{p}$	maximum positive deviation of service times
$\mathcal{R}_d$	set of all routes on day d
R	revenue
$\underline{T}$	earliest availability of patients
$\bar{T}$	latest availability of patients
$\tilde{t}$	expected travel times
$\hat{t}$	maximum positive deviation of travel times
$\mathcal{U}^p$	set of service time scenarios
$\mathcal{U}^t$	set of travel time scenarios
u^*	optimal second-level dual solution
v_i	required number of visits for patient i
$\bar{v}$	upper bound on value per visit
v^*	optimal first-level dual solution
$\underline{W}$	earliest availability of caregivers
$\bar{W}$	latest availability of caregivers
w^*	optimal first-level dual solution
y^*	optimal second-level dual solution
z^*	optimal second-level dual solution
Γ^p	uncertainty budget for service times
Γ^t	uncertainty budget for travel times