

The Impact of Waiting Time and Treatment Modality: An Empirical Analysis of a Pediatric Speech-Language Therapy Program

by

Benjamin Ravenscroft

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Management Sciences

Waterloo, Ontario, Canada, 2025

© Benjamin Ravenscroft 2025

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Demand for multi-appointment, outpatient care services is high and growing, pressuring overburdened healthcare systems and highlighting the need for operational improvements. In this study, we investigate operational and scheduling-related factors, and their impact on the number of appointments patients require to be discharged, as well as their appointment attendance behaviour. We first utilize data from over 1,700 pediatric speech-language therapy patients to study the associations between waiting time for service, treatment modality, and the number of appointments patients require. Using a continuation ratio model, we estimate that, on average, each additional year a patient waits to begin treatment is associated with a 26.3% increase in the total number of appointments needed. Furthermore, we show that offering all appointments in-person, versus via telemedicine, patients require 8% fewer total visits to achieve their desired outcomes. While these are the average effects across the patient population, we also find significant differences between patients of different severities and the relationships between waiting time, modality, and the number of appointments needed. Secondly, we use a mixed-effects multinomial logistic regression to investigate the relationship between appointment modality and appointment attendance. Although in-person treatment is associated with patients requiring fewer total appointments to achieve desired outcomes, in-person appointments are also 4.42% less likely to be attended. This highlights the need for management to consider the effects of scheduling decisions on multiple drivers of system throughput.

Acknowledgements

My sincerest thanks to my supervisor, Prof. Hossein Abouee Mehrizi, for his support, guidance, and mentorship. Working with him has not only made me a better researcher but also given me valuable insights into teaching and opened avenues to future academic opportunities.

I also want to extend my thanks to Brendan Wylie-Toal, the whole of the research and innovation team at KidsAbility, and their amazing management team. Without them this project would not have been possible, and I hope this work has the positive impacts on pediatric rehabilitation we hope it does.

Finally, I would like to thank my friends and family for their love and support.

Table of Contents

Author's Declaration	ii
Abstract	iii
Acknowledgements	iv
List of Figures	vii
List of Tables	viii
1 Introduction	1
2 Literature Review	5
3 Empirical Setting	8
4 Background and Hypotheses Development	10
4.1 Direct Determinants of the Number of Appointments Needed	10
4.1.1 Referral to Assessment Waiting Time	10
4.1.2 The Effects of Treatment Modality	12
4.2 Cancellation Behavior	13
4.3 Explanatory Variables	17

5	Empirical Strategy	20
5.1	Number of Appointments Needed	20
5.2	Number of Appointments Needed Cont'd	23
5.3	Cancellation and No-Show Probabilities	23
6	Empirical Results	25
6.1	Number of Appointments Needed	25
6.2	Cancellation and No-Show Probabilities	29
7	Counterfactual Analysis	33
7.1	Simulation Design	33
7.2	Effects of Wait Time	34
7.3	Effects of Modality	35
8	Conclusion	38
	References	41
	APPENDICES	45
A	Additional Figures and Tables	46

List of Figures

6.1	Estimation of the Effect of Wait Time and Modality on Restricted Mean Survival Time (RMST)	27
6.2	Estimation of the Effect of Wait Time and Modality on Restricted Mean Survival Time (RMST), Stratified by Severity Level	29
7.1	Best Performing Strict Priority Policies Based on Class Costs for Long-Run Age-Out Rates and Waiting Times (Class L Cost Normalized to 1)	35
A.1	Relationship Between Lead Time and the Number of Appointments Scheduled Between When the Focal Appointment was Booked and When it Was Scheduled to Occur.	46

List of Tables

4.1	Summary Statistics of Key Dependent and Control Variables	19
6.1	Survival Analysis Model Results	26
6.2	Survival Analysis Model Results Stratified by Severity Level	28
6.3	Multinomial Logistic Model Results	30
6.4	Multinomial Logistic Model Results - In-Person Only	32
7.1	Arrival Rates, Baseline Service Rates, and Wait Time Effects for Three Severity Classes	34
7.2	Simulation Results for Different Modality Policies	36
A.1	Standardized Mean Difference of Covariates Between In-Person and Virtual Appointments	47
A.2	Multinomial Logistic Model Results - Average Marginal Effects	47
A.3	Multinomial Logistic Model Results - Average Marginal Effects (In-Person Appointments)	48
A.4	Simulation Results for Different Modality Policies - $\rho = 0.97$	48

Chapter 1

Introduction

Many healthcare services, such as speech-language therapy (SLT), rehabilitation therapy, and cancer treatment, require multiple appointments over extended periods. The demand for these services is high and continues to grow, placing immense pressure on already overburdened healthcare systems and highlighting the urgent need for operational improvements to enhance system throughput. For example, approximately 7.7% of children aged 3 – 17 in the United States have communicative disorders, with 55% of these cases receiving intervention through SLT [?]. This equates to nearly 21 million children annually. However, children are not the only demographic requiring rehabilitation therapy; SLT also plays a crucial role in managing conditions such as Alzheimer’s, dementia, and stroke, all of which are becoming increasingly prevalent as populations age [?]. Similarly, other multi-appointment healthcare services, such as occupational or behavior therapy for autism spectrum disorder, the prevalence of which has surged by 400%¹ since the early 2000s, face rising demand.

In this paper, we focus on pediatric SLT, specifically a publicly funded, preschool program operating in Ontario, Canada, serving children typically aged 0 – 4 years. Such programs provide treatment for children with speech-language disorders (SLDs), which encompass difficulties producing sounds (i.e., speech disorders), understanding or combining words to express ideas (i.e., language disorders), and swallowing. Due to the detrimental effect of SLDs on learning and social skill development, early intervention is imperative. However, the waiting times for SLT services, particularly publicly funded programs such as those delivered by Children’s Treatment Centres (CTCs) in Ontario, are high and rising. This has been driven primarily by demand outpacing supply growth. For instance, some

¹<https://www.cdc.gov/autism/data-research/index.html>

CTCs have reported annual SLT services referrals rising 75% between 2019 and 2024, which contributed to associated wait time increases of 13 – 14% annually. Clinician shortages are expected to continue in Canada, as the number of speech-language pathologists (SLPs) entering the workforce is projected to fall short of the demand increase through 2031 [?]. The issue of SLP shortages is not isolated to Canada, as Australia and the United Kingdom face similar challenges, nor is it isolated to universal healthcare, as the United States faces related issues [19, 30].

Motivated by the need to identify methods to reduce CTCs’ waiting times independent of increasing staffing levels, our research aims to empirically investigate operational and scheduling factors related to throughput. Preschool SLT differs from healthcare settings more commonly examined in operations management literature in that it is a) a multi-appointment, outpatient care setting, and b) compared to healthcare services such as cancer care, treatment planning decisions largely rest in the hands of individual clinicians and are not algorithmically guided. The former aspect of SLT means that results and insights from single-appointment healthcare settings may not necessarily hold in the setting of SLT. The latter aspect emphasizes the need to estimate the treatment process from empirical data, as we lack tightly bounded clinical guidelines that can be relied upon for understanding treatment dynamics. Our study focuses primarily on the two determinants of throughput in multi-appointment settings: 1) the number of appointments patients need, and 2) the attendance behavior of patients. We aim to investigate the relationship between initial waiting time and modality of appointments (i.e., in-person vs. virtual), and these determinants.

In the context of single-appointment healthcare settings, such as the intensive care unit (ICU) setting studied by [?], delays in receiving care have been shown to sometimes result in adverse outcomes and longer length of stay. It is important to note that these settings are 1) single-appointment care settings, and 2) the longer stays resulting from delays are likely attributable to patient conditions deteriorating while waiting for intervention. Our setting is different in two ways. First, we investigate a multi-appointment, outpatient care setting. Second, if waiting times for treatment are associated with SLP patients needing more appointments, there is reason to believe the driving mechanism is not the deterioration of the patient’s condition but rather their learning curve flattening as waiting time grows. Understanding the relationship between waiting time and the number of appointments patients need is critical when considering approaches to improve throughput, such as implementing priority-ordered waitlists. It is particularly important to understand whether different patient types respond to waiting times differently.

Like many non-emergency healthcare organizations, SLP providers began to increasingly adopt telehealth during the pandemic. While telehealth can make access to care more

convenient by eliminating geographic and time constraints for patients and providers, its clinical and service efficiency is relatively unknown in pediatric SLP and multi-appointment, outpatient, chronic care more broadly (excluding mental health treatment). From an organizational perspective, understanding how telehealth influences factors of throughput, such as the number of appointments patients require and their attendance behavior, is critical when allocating capacity and making scheduling decisions. Furthermore, as [21] report, the majority of patients still prefer in-person visits across a variety of healthcare conditions. Were organizations to have evidence of the clinical and service efficiency of telehealth, they may be able to address the common belief that diagnosis and treatment are more accurate and effective with in-person care, consequently increasing patient acceptance of telehealth.

Findings and Contributions. Our paper makes several important contributions to the literature on multi-appointment service systems, particularly rehabilitation systems. To our knowledge, this is the first paper to analyze the relationship between waiting time and the number of appointments patients require, as well as the joint relationship between the use of telehealth, the number of appointments needed, and appointment attendance behavior. Specifically, we make the following contributions:

1. Using retrospective data from over 1,700 patients and 15,800 appointments, we empirically estimate the impact of initial waiting time and appointment modality on the number of appointments patients need to achieve desired outcomes. We use a discrete-time survival analysis approach with inverse proportional censoring weighting to account for right-censored data with a known censoring mechanism. Firstly, we find that for medium- and high-severity patients, a one-year delay is associated with a 9 – 21% increase in the number of appointments needed. However, there is no evidence of a significant effect of waiting time on the number of appointments needed by low-severity patients. Secondly, we find that for medium- and high-severity classes receiving the majority of appointments virtually significantly increases the number of appointments needed compared to receiving the majority in-person. For low-severity patients, there is no significant evidence that modality affects service efficiency.
2. By analyzing granular, historical scheduling data for over 19,000 appointments, we find that virtual appointments (i.e., telehealth) are 4.42% more likely to be attended than in-person appointments. Interestingly this is driven by a significantly lower likelihood of no-showing or canceling an appointment within 24 hours of the start time. Furthermore, this observation is uniform across patient severity levels.
3. Using simulation results, we demonstrate that ignoring these relationships can lead

to significant negative impacts on throughput and, consequently, greatly increase waiting times. Firstly, we show that in the presence of the waiting time effect on the number of appointments needed, a strict priority policy can significantly outperform the currently used first-come-first-serve (FCFS) policy concerning throughput. Importantly, the optimal strict priority policy does not necessarily follow the $c\mu$ -rule. Secondly, we consider static scheduling policies in which each patient category is designated a fraction of appointments to receive in-person versus virtually. We find that it is important to consider both the effect of modality on the number of appointments needed as well as appointment attendance behavior. Doing so can significantly improve throughput compared to the organization’s current policy.

Chapter 2

Literature Review

This paper relates to four streams of literature: empirical work investigating the effect of waiting time and modality on service durations, empirical investigations of scheduling factors associated with appointment attendance behavior, queueing models with queueing time-dependent service durations, and appointment scheduling in healthcare.

Significant effort has been made to research the impact of delays on patient outcomes, such as hospital length of stay (LOS). For example, [5] find that the median hospital length of stay and in-hospital mortality rates are significantly higher for patients whose ED to ICU transfer was delayed by more than 6 hours. The main drawback of this and similar studies in the medical literature is their treatment of delays as a binary outcome. On the other hand, [6] treat emergency department (ED) to ICU delays as continuous. They find that one additional hour of delay is associated with an 11.37 – 38.21% increase in ICU LOS for select patient categories, like renal and vascular patients. Furthermore, these studies consider hospital settings where service time is defined in terms of a single LOS. We verify, in the context of a multi-appointment SLT setting, that delays between treatment referral and assessment are associated with patients needing more appointments. Moreover, this association is not uniform across patient severity levels.

With the rise of telehealth usage, there is a growing body of literature investigating the effectiveness of telehealth on the likelihood of patients needing follow-up or duplicate appointments. In the context of acute respiratory illnesses, [29] find that telehealth diagnosis appointments resulted in a 5 times higher likelihood of needing an in-person follow-up compared to an in-person diagnosis appointment. Contrarily, [12] found that follow-up rates were the same for virtual and in-person appointments in the context of acute, non-urgent care. One limitation of these studies is they consider only the likelihood of needing one

more follow-up appointment due to their focus on acute, rather than chronic, care. [1] show that adoption of secure messaging, or “e-visits”, by primary care providers increase office visits by 6% and decrease the number of new patients providers see each month by 15%; however, they do not significantly impact health outcomes. We extend these findings to a multi-appointment care setting, demonstrating that for some patient categories, increased usage of telehealth methods is associated with an increase in the number of appointments needed to achieve adequate outcomes.

Our work leverages electronic health records and scheduling data to parametrically estimate the associations between waiting time, treatment modality, and the number of appointments patients need, which is an analog of service time in a multi-appointment setting. We use a discrete-time survival analysis approach to address the estimation, similar to the methods used by [15], to examine the relationship between medication adherence and readmission.

Significant effort has been invested into researching the reasons patients might miss appointments. To date, the works of [8] and [2] remain some of the most comprehensive literature reviews on the topic. They cite patient demographic features, patient-provider interactions, previous appointment-keeping history, and appointment lead time as predictors of no-show behavior. One of the primary limitations of the existing literature is the common grouping of no-shows and cancellations or exclusion of cancellations from the analysis. [22] attempt to address this by distinguishing three appointment outcomes: attendance, no-show, and cancellation. The main predictors they identify are lead time, financial payer, patient age, and prior attendance history. Recently, the medical literature has begun to investigate the association between telemedicine and appointment attendance behavior, finding mixed results. For example, while [23] estimated the no-show rate for telemedicine appointments (12%) is half that of in-person appointments (25%), [3] estimated the odds of a no-show for a telemedicine appointment is 42% higher than an in-person appointment. We add to this body of literature in several ways. Firstly, we consider four appointment outcomes: 1) attendance, 2) no-show, 3) cancellation within 24 hours of the appointment (last-minute cancellations), and 4) cancellations greater than 24 hours before the appointment (advance cancellations). Secondly, we not only investigate the relationship between telemedicine and attendance behavior, but also the relationships between the number of future known bookings, patient drive time to their clinic, and attendance behavior.

Motivated by our empirical findings, we consider simple static scheduling policies that can improve overall system throughput and other patient outcome measures in our setting. [9] investigate shortest-job-first (SJF) and shortest-remaining-processing-time (SRPT) scheduling policies in the setting of a $M/GI/s+GI$ queue. They show that SJF/SRPT is asymptotically indistinguishable from a carefully designed two-class priority

rule. [28] considers appointment scheduling with a bulk arrival and two patient classes who have different service time responses to waiting time. They find that the best scheduling policy depends on the parameterization of the initial service times and the responses to waiting time. In our work, we extend the concept of a strict-priority waitlist to a setting where service times for different patient classes are differentially dependent upon their wait times. Furthermore, we provide evidence that in some settings, SJF (based upon the baseline service time) is not the best priority ordering.

Our second counterfactual analysis is connected to works such as [34], [24], and [17] that investigate the impact of telehealth on system throughput when their effect on the need for duplicate or follow-up visits varies. [34] and [17] find that the use of telehealth and in-person appointments need to be carefully balanced, dependent on the efficiency of telehealth visits. Too much prioritization and telehealth capacity can lead to providers delivering more appointments while simultaneously seeing fewer patients. Key aspects of our work include appointment modality not only modifying the likelihood of needing more appointments but also the appointment attendance behavior. We show that these factors should be considered simultaneously by managers, as telehealth may simultaneously increase the number of appointments a patient needs while also decreasing their odds of missing or canceling appointments.

Chapter 3

Empirical Setting

In Chapter 3 we elaborate on the clinical setting, followed by the patient treatment timeline, and concludes with a data description. We collected data from 6,834 patients across 7 clinics, who received SLT between December 2020 and October 2023. Our data set provides us with a rich set of variables related to the referral, assessment, and discharge processes for patients, demographic and disorder severity information, and scheduling information.

We exclude patients with missing data; specifically, we remove patients with no recorded referral date or severity score, and patients who canceled all of their appointments. A large component of this reduction is the exclusion of patients who were referred to the program before data collection began in December 2020, but had yet to begin or were currently undergoing treatment when data collection began. Furthermore, we only include patients whose treatment began and ended during the data collection period. This yields a sample size of 1,751 patients and 15,806 attended appointments suitable for use in the analysis of factors driving the number of appointments patients need. Comparatively, we include the full 6,834 patient sample in our analysis of predictors of appointment-keeping behavior. There were 47,460 appointments booked across these patients, 36,971 (77.90%) of which were attended.

A patient's treatment pathway begins with a referral to the CTC, typically made by a pediatrician, family physician, or a SLP at a community outreach event. The referring clinician provides their perspective on the patient's type and severity of disorder, which Clinical Scheduling Assistants (CSAs) use to assign the patient a priority code. The patient is subsequently added to the program waitlist which functions as a FCFS waitlist.

When a clinician is ready to accept a new patient onto their caseload, they contact a CSA who pulls a patient from the waitlist. Generally, the CSA picks a patient who is

closest to the front of the waitlist in conjunction with the service needed, the patient's treatment location preference (as there are multiple treatment sites), and currently available treatment resources. We note that, in practice scheduling deviates slightly from a FCFS policy during times of the year when the clinic needs to prioritize patients on the waitlist about to age-out of service eligibility.

Once assigned a clinician, the patient begins service with an assessment appointment. During this appointment, the clinician assesses the client's speech-language disorder, the family's desired treatment outcomes, and possible barriers to treatment, such as an unstable home environment. The clinician subsequently outlines a set of treatment goals for the patient and the treatment plan, including the estimated number of treatments required to reach the stated goals. Furthermore, they provide the family with a home treatment program to follow, which includes exercises for the patient to perform between appointments. Patients subsequently receive a series of appointments that conclude in one of two outcomes. 35% of patients are discharged for achieving adequate outcomes, whereas 65% of patients are discharged due to entering school, which renders them ineligible for treatment in the preschool program. These patients are subsequently referred to school-based rehabilitation services (SBRS), where they join another waitlist. This can significantly disrupt a child's rehabilitation process as SBRS programs generally have waiting times in the order of several years in Ontario.

In addition to the number of appointments patients receive and the reason for their discharge, our data also enable us to construct multiple measures related to appointment-keeping behavior. Firstly, we know which appointments were canceled due to a patient cancellation, clinician cancellation, or no-show. Secondly, we know the date on which each appointment was canceled, which allows us to compute the advance notice provided at the time of each cancellation. We classify patient cancellations into three categories: no-show, last-minute cancellation, and advance cancellation. No-shows, last-minute cancellations, and advance cancellations correspond to 4.52%, 6.96%, and 4.63% of all appointment bookings, respectively.

In the next chapter we describe our empirical hypotheses in detail and provide a more detailed description of the data we employ to test these hypotheses.

Chapter 4

Background and Hypotheses Development

The determinants of system throughput in our setting are complex and interconnected. For example, the number of appointments patients require to achieve appropriate outcomes directly impacts system throughput, while the time it takes to deliver these appointments is partially affected by the frequency with which the patient cancels or misses their bookings. In this section, we present multiple hypotheses and explain their direct and indirect impacts on throughput.

4.1 Direct Determinants of the Number of Appointments Needed

4.1.1 Referral to Assessment Waiting Time

We begin by considering the effect of the waiting time a patient experiences before beginning treatment on the number of appointments they receive. From a managerial perspective understanding this relationship is critical. If the relationship is positive, it can create a positive feedback loop that detrimentally affects system throughput. Furthermore, if the effects vary by patient type, it may affect the optimal choice of waitlist prioritization policy.

In SLT, the wait-and-see approach was once believed to be beneficial for toddlers who initially failed a language screening, supported by the theory that they may experience

spontaneous improvement while waiting for treatment. However, [4] provides evidence that this approach is outdated and is not supported by recent studies. However, when waitlists are long, early identification and referral for treatment does not necessarily translate to early intervention. Waiting times at CTCs have grown steadily over the last decade, as demand has consistently outpaced clinical supply of SLT. This indicates low elasticity of demand for publicly-funded treatment with respect to waiting times, similar to the findings of [18] with respect to surgical wait times, and emphasizes the need to better understand the relationship between waiting times and throughput when deciding how to manage waitlists.

The literature on the dependency of service times on waiting times is disparate, equivocal, and refers mainly to settings where service is delivered in a single, rather than multiple, appointment(s). For example, [6] show that delays in the admission of patients from the ED to the ICU can result in longer stays, however, this is only the case for 3 of the 9 patient categories investigated. Furthermore, [6] study a single visit setting. [14] investigated changes in disorder severity of pediatric patients referred for SLT while they waited to begin service. Although patients made an average improvement of 0.5 on the Therapy Outcomes Measures Impairment Scale (TOM-I), the majority of patients were still deemed to need therapy after reaching the front of the waitlist. Furthermore, waiting times were found to be insignificantly correlated with changes in disorder severity in both univariable and multivariable regressions. Finally, although [14] study the effect of waiting on disorder severity and the *binary* need for treatment in our setting, they do not investigate the effect of wait times on the number of treatment appointments needed or treatment outcomes.

Based on findings from single appointment settings and the existing literature on the effect of waiting times on speech-language disorder progression, we postulate the following alternate hypotheses:

Hypothesis 1A *Longer waiting times are associated with patients needing more treatment appointments to be discharged for achieving desired outcomes.*

Hypothesis 1B *The number of appointments patients need to be discharged for appropriate outcomes has an inverted U-shaped response to their initial waiting time.*

In Hypothesis 1A, we assert that as a patient waits longer for treatment, the more their learning curve flattens. This results in patients requiring more treatment to achieve the desired outcomes.

Hypothesis 1B is motivated by the observation in [14] that some patients with SLDs spontaneously improve while waiting for treatment. We hypothesize that as patients wait

for service, their learning curve flattens, requiring more treatments to achieve the desired outcomes; however, patients who wait a very long time are likely to spontaneously improve, consequently decreasing the amount of treatment required to achieve the desired outcomes. This leads to an inverted U-shaped relationship between wait time and the number of appointments needed.

4.1.2 The Effects of Treatment Modality

Telehealth, defined here as service delivery via video conferencing, telephone, or email, saw a surge in usage during the COVID-19 pandemic. Although many healthcare settings returned to in-person care as restrictions eased, telehealth remains valuable in some contexts, particularly for reducing indirect care costs, such as reducing travel time for patients. For instance, [32] reported high patient satisfaction with telehealth visits in an SLT setting, and significant patient interest in its continued availability.

The efficacy of telehealth in adult SLP is relatively well-established (see [10]); however, the literature for pediatric patients, particularly preschool-aged patients, is limited. Although [33] demonstrate that telehealth is efficacious in terms of patient conditions improving—for instance, Functional Communication Measures improve significantly when patients are treated using telehealth—there is little existing research comparing the efficiency of telehealth and in-person treatment in an SLT context.

This study focuses on the efficiency of different treatment modalities, measured by their effect on the number of appointments patients require to be discharged with desired outcomes. Assuming equivalent efficacy between modalities, differences in efficiency can highlight disparities in resource utilization. For instance, if virtual treatment requires 15 appointments to discharge a patient, but in-person care only requires 10, in-person treatment is more efficient.

Additionally, we are interested in examining how efficiency varies with patient-level factors, such as disorder severity. Some disorders, such as swallowing and voice disorders, require physical intervention unsuitable for telehealth. Others, such as stuttering, require clinicians to be able to accurately assess aspects of voice, like quality, pitch, and hoarseness. [11] show these factors can be hard to assess accurately in a telehealth setting due to distortion from audio recording devices. These limitations indicate that the efficacy and efficiency of telehealth may depend on the clinical needs of the patient. Misalignment between treatment modality and patient needs may lead to inefficiencies, such as unnecessary follow-up appointments, which reduce the overall benefits of telemedicine.

Based on the current body of literature, we form the following hypotheses:

Hypothesis 2A *In-person treatment is more efficient compared to telehealth treatment, resulting in patients requiring fewer appointments to achieve desired outcomes.*

Hypothesis 2B *The difference in efficiency between in-person treatment and telehealth treatment is greater for patients with high, compared to low, disorder severity.*

Hypothesis 2C *Mixing treatment modalities is less efficient than using a uni-modular treatment style, due to frequent changes in the treatment environment disrupting patient learning in a multi-appointment setting.*

4.2 Cancellation Behavior

Although the number of treatments patients need to receive is the primary driver of system throughput, the number of bookings necessary to deliver these treatments is a secondary driver. In a multi-appointment setting where patients can cancel or miss appointments, the number of bookings and the number of appointments delivered are not necessarily equal. For example, a patient that requires six appointments and cancels 20% of their scheduled appointments will use up seven to eight booking slots. This has negative consequences on the throughput when these ‘unattended’ appointments are no-shows or last-minute cancellations as these appointment slots are hard (or impossible) to fill in a setting without overbooking or walk-in appointments.

In many service settings, walk-in appointments and overbooking are used as strategies to mitigate the negative effects of no-shows after they occur. In the case of SLT, these are highly challenging strategies to implement. As service is provided across multiple appointments that must be delivered by the same clinician, ideally with a specific frequency, walk-in appointments are difficult to accommodate. Furthermore, overbooking has a much higher cost potential than in other service settings for several reasons. First, when treating preschool-aged children, this strategy requires service providers to offer childcare services for overbooked patients. Secondly, overbooked patients can only fill no-shows and cancellations for their clinician. This greatly increases the probability of these patients being treated in overtime, which increases costs for service providers and can be impossible in cases where clinician contracts prohibit overtime work.

For the aforementioned reasons, many of the strategies suggested in the existing literature for mitigating the negative impact of appointment non-attendances do not apply in our setting. To limit the negative effects, strategies that reduce the occurrence of no-shows

and cancellations, particularly last-minute cancellations, must instead be employed. One such strategy is that of optimizing the booking window when the likelihood of no-shows increases monotonically with appointment lead time. This approach was proposed theoretically by [16] and shown to have significant benefits in an SLT setting in [25]. In this work, we analyze other facets of appointment attendance behavior with a goal of identifying factors that can be leveraged to reduce the rates of no-shows and cancellations.

Primarily, we are interested in the relationship between appointment modality and the number of known future bookings, and the likelihood of a last-minute cancellation or no-show. In Section 4.1, we hypothesized that appointment modality affects the number of treatments patients require to achieve appropriate outcomes. However, the amount of bookings it takes to deliver these treatments (i.e., attended plus canceled appointments) is dependent on the non-attendance rate, making the choice of the optimal appointment modality non-trivial. For example, if 100% in-person treatment translates to an expectation of a patient needing 15 appointments, but in-person appointments are canceled 20% of the time, they will need an expected 18 bookings to receive 15 treatments. However, if 100% virtual treatment translates to a need for 16 appointments, but virtual appointments are canceled 10% of the time, the patient will only need an expected 17.6 bookings to achieve appropriate outcomes. Secondly, if the likelihood of cancellation or no-show increases as the number of appointments delivered increases, patients who require more treatment will also have more cancellations, further exacerbating their use of treatment capacity.

It is natural to think that telemedicine appointments would have lower no-show and last-minute cancellation rates compared to in-person appointments, as they reduce the logistical needs for attending an appointment. For example, travel to the treatment site is no longer required, which may make attending an appointment more feasible for an individual with a busy schedule, and reduce the number of cancellations and no-shows due to adverse road conditions.

Based on the existing literature, we form the following hypothesis:

Hypothesis 3 *Telemedicine appointments (i.e., telephone and virtual appointments) are less likely to incur cancellations or no-shows compared to in-person appointments.*

Regarding the association between known future bookings and patient attendance behavior, we form the following two hypotheses:

Hypothesis 4A *The likelihood of a no-show decreases as a patient has more known future bookings after their upcoming appointment.*

Hypothesis 4B *The likelihood of a patient canceling their upcoming appointment in advance (i.e., greater than 24 hours in advance) grows with the number of known future bookings.*

In the case of the organization being studied, there is a policy that aims to reduce no-show rates by putting patients who no-show twice without valid cause back on the treatment waitlist. Analysis of the available data shows that this policy is strongly enforced, with less than 2% of patients having more than 2 recorded no-shows. We argue that the more future known bookings a patient has, the greater the incentive is for them to not no-show, as there is a significant known cost in the event they are put back on the waitlist. This rational underpins Hypothesis 4A.

In the case of Hypothesis 4B, we argue that the likelihood of an advance cancellation occurring increases with the number of future known bookings for two reasons. First, there is no penalty for an advanced cancellation, unlike a no-show. Secondly, if a patient has more known future bookings, their perceived loss from canceling their upcoming appointment is lower.

The final hypothesis of interest is the relationship between drive time and appointment attendance behavior. Given the outpatient setting, patients with in-person appointments need to travel, generally from their homes, to their respective clinic for treatment. We focus on the estimated drive time from patients' homes to their respective clinic, as opposed to straight line distance, due to the high number of highways/freeways as well as residential roadways in the geographic area of the clinic network. As these paths of travel flow at significantly different speeds (e.g., 100km/h versus 40km/h) we assert that it is possible for equal euclidean distances to have different driving times. Logically, what is likely to cause a no-show or last-minute cancellation due to travel is that the patient is unable to arrive at the clinic on time for their appointment. This outcome is more directly a function of the time it will take for them to drive from their home to the clinic as opposed to the absolute distance from their home to the clinic.

The existing literature on travel-related predictors of appointment attendance behavior in outpatient care settings lies primarily in the medical literature and typically examines the association between distance and no-show behavior. Few studies have examined the effect of driving time or the association between distance/driving time and cancellation behavior, in addition to no-show behavior. Furthermore, while some studies elucidate a positive association between distance and no-show likelihood, such as [13], others, such as [20], note a negative correlation. Although we are interested in the association between drive time and no-show behavior, we are more focused on understanding the complex relationships

and interplay between drive time and attendance, no-shows, and cancellations, as these are not independent outcomes and have different managerial impacts.

Based on the existing literature, there appears to be a complex interplay between drive time and no-show and advanced cancellation behavior. This leads to the construction of the following two related hypotheses:

Hypothesis 5A *The likelihood of a no-show decreases as driving time from the client location to their clinic increases.*

Hypothesis 5B *The likelihood of an advanced cancellation increases as driving time from the client location to their clinic increases.*

One cause for no-shows and cancellations is scheduling conflicts on the patient's end. Consider the case of a minor scheduling conflict, for example, the patient's parent has a meeting scheduled to end 15 minutes before their appointment. If they live close to the clinic, for example, a 10-minute drive away, the parent may still believe they have a high probability of making it to the appointment on time. Thus, they choose not to cancel their appointment in advance, instead taking the risk of no-showing, leading to a higher no-show likelihood and a lower likelihood of them informing the clinic of the scheduling conflict in advance and canceling/rescheduling the appointment. Conversely, suppose this family lives 30 minutes away from the clinic, and the parents' meeting is scheduled to end 15 minutes before the start of their child's appointment. In that case, they are more likely to cancel/reschedule the appointment in advance as they are certain they will not be able to make it to the appointment on time due to the conflict. Therefore, as drive time increases, the likelihood of an advance cancellation rises and the likelihood of a no-show decreases.

If these hypotheses hold they have important managerial implications, regardless of whether appointment attendance likelihood is associated with drive time. As previously mentioned, for the organization being investigated, advanced cancellations carry a much lower cost in terms of wasted treatment capacity compared to no-shows. Consequently, given the choice, if the likelihood of attendance itself cannot be changed, it is preferable to enact policies that shift non-attendances to higher rates of advanced cancellations and lower rates of no-shows, compared to the other way around.

4.3 Explanatory Variables

Our detailed scheduling data allows us to construct patient treatment timelines and multiple explanatory variables. To be specific, for each patient i , we know the initial waiting time (i.e., the time from referral to assessment), denoted by $Wait_i$, the severity level, denoted by Sev_i , the age at referral, denoted by $RefAge_i$, the treating clinician denoted, by $Clinician_i$, and the location where all *in-person* treatments were received, denoted by $Office_i$. Of the 1,751 patients used in the analysis of the number of appointments needed, 21% are low severity, 54% are medium severity, and 25% are high severity.

Our data also allows us to construct several time-varying explanatory variables. To be specific, we can calculate the time elapsed between consecutive appointments, and the cumulative average interval between consecutive appointments from appointment 1 to j , denoted by Gap_{ij} for patient i . This variable allows us to estimate the impact of appointment frequency on patient outcomes. Additionally, appointments can be delivered in-person or via telehealth methods. Using the data for the modality of delivery of each appointment, we can calculate the fraction of appointments that were delivered in-person or through telehealth. Specifically, we can calculate the cumulative fraction of appointments $1, \dots, j$ that were delivered in-person for patient i , denoted by $Face_{ij}$.

We are also able to construct multiple explanatory variables related to patient appointment attendance behavior. First, we know the date on which each appointment was scheduled. We denote the date on which appointment j of patient i was scheduled by T_{ij}^{sched} , and the date for which it was scheduled to occur by T_{ij}^{appt} . Using these variables, we can compute the lead time for each appointment booking, denoted by $LeadTime_{ij} = T_{ij}^{appt} - T_{ij}^{sched}$. Furthermore, we can construct measures of the historical attendance behavior for each patient. Specifically, we can calculate the number of appointments patient i had attended, canceled, and no-showed before time t , denoted by $PrevAppt_{it}$, $PrevCanc_{it}$, and $PrevNS_{it}$, respectively. We can also calculate the number of future appointments patient i had scheduled before time t , denoted by $FutureBookings_{it} = \sum_{j=1}^{m_i} \mathbb{1}(T_{ij}^{sched} \leq t < T_{ij}^{appt})$, where m_i denotes the total number of appointments scheduled for patient i in the data.

Furthermore, our data provides other appointment scheduling-related information, such as the time of day for each appointment and its modality (i.e., in-person or virtual). We use publicly available data sources to extract additional relevant information such as historical weather data (temperature, snowfall, and rainfall). Moreover, using the postal code of a patient’s home address and their clinician’s office, we calculate the approximate travel distance and travel time from the patient’s home address to their clinician’s office. Table 4.1 provides the summary statistics of patient-level and appointment-level covariates related

to the number of appointments needed and appointment attendance behaviour.

Table 4.1: Summary Statistics of Key Dependent and Control Variables

Variable	Observations	Mean	St. Dev.
Baseline Patient Covariates			
Number of Treatment Visits	1,751	9.03	5.01
Event (1 = Discharged, 0 = Aged-out)	1,751	0.35	0.48
Age at Referral (years)	1,751	1.93	0.57
Waiting Time (years)	1,751	0.91	0.42
Time-Varying Patient Covariates			
Cumulative Percentage of Appointments In-Person	15,806	0.45	0.35
Interval between Appointments (days)	15,806	27.27	23.67
Appointment Attendance Distribution (Patient-Level)			
Attendance (%)	22,169	84.87	19.80
No-Show (%)	22,169	4.85	13.06
Last-Minute Cancellation (%)	22,169	5.99	11.28
Advance Cancellation (%)	22,169	4.28	10.12
Attendance Behavior Covariates			
Lead Time (days)	22,169	30.30	21.95
Number of Previous Cancellations	22,169	1.06	1.54
Travel Distance (km)	22,169	9.43	6.21
Travel Time (min)	22,169	12.31	5.17
Temperature (°C)	22,169	9.44	9.01
Daily Rainfall (mm)	22,169	1.91	4.39
Daily Snowfall (mm)	22,169	0.28	1.04

Chapter 5

Empirical Strategy

In Chapter 5 we describe the empirical strategy we employ to estimate the impact of the key variables of interest on the number of appointments needed and the cancellation behavior. In what follows, we attach a subscript i to the various variables of interest to denote the patient, and a subscript j to denote the appointment number.

5.1 Number of Appointments Needed

To test Hypotheses 1-2, we need to estimate the impact of waiting time and treatment modality on the number of appointments patients need to achieve the desired outcome. However, we only observe the number of appointments patients received and whether they were discharged for having reached the desired outcome. Let N_i denote the number of appointments patient i needed, and let $\tilde{N}_i$ denote the number of appointments patient i received in the data. We cannot use a generalized linear model such as a Poisson or a negative binomial regression because our observations are censored, that is, we do not observe the total number of appointments a patient needed if the patient did not reach the desired outcome before aging out. If we were to ignore this concern, the parameter estimates could be incorrectly estimated as the number of appointments received is an under-representation of the number of appointments censored patients needed. To rectify this concern, we use survival analysis as it accounts for the censoring process. Survival analysis is frequently used in the medical and reliability literature where the time-to-event of a deteriorating system is of interest. It has also been used to model credit risk (particularly, time to default) for loans and bonds as well as inter-purchase times; for example, [31] and [27].

The Survival Model.

Let N_i denote the number of appointments patient i received, and δ_i be a binary variable that equals one if patient i reached the desired outcome (before aging out) and zero otherwise. Let X_i^b denote the baseline covariates for patient i , and $X_i^t = (X_{ij}^t)$ denote the time-varying (appointment specific) covariates for patient i , where X_{ij}^t denotes the time-varying covariates at the time of appointment j for $j = 1, \dots, N_i$. For each patient i for $i = 1, \dots, m$, we observe $(N_i, \delta_i, X_i^b, X_i^t)$. An equivalent representation of (N_i, δ_i) is $(\delta_{i1}, \dots, \delta_{iN_i})$, where δ_{ij} for $j = 1, \dots, N_i$ is the binary variable indicating whether patient i was discharged after appointment j for having achieved the desired outcome, defined as follows:

$$\delta_{ij} = \begin{cases} 1 & \text{if } j = N_i \text{ and } \delta_i = 1 \\ 0 & \text{otherwise.} \end{cases}$$

The key quantity of interest is the hazard, that is, the instantaneous risk, of the patient reaching the desired outcome, which is defined as follows:

$$\lambda_{ij}(X_i^b, X_{i,j}^t) = \mathbb{P}(\delta_{ij} = 1 \mid \delta_{i,j-1} = 0, X_i^b, X_{i,j}^t).$$

We employ the *continuation ratio model* proposed by [7], in which the hazards, λ_{ij} , are parameterized as follows:

$$\log\left(\frac{\lambda_{ij}}{1 - \lambda_{ij}} \mid X_i^b, X_{i,j}^t\right) = \alpha_j + \beta X_i^b + \gamma X_{i,j}^t, \quad (5.1)$$

where α , β , and γ are given parameters. Equation (5.1) can be rewritten as follows:

$$\lambda_{ij} = \frac{\exp(\alpha_j + \beta X_i^b + \gamma X_{i,j}^t)}{1 + \exp(\alpha_j + \beta X_i^b + \gamma X_{i,j}^t)}.$$

Then, the likelihood can thus be expressed as follows:

$$L = \prod_{i=1}^m \prod_{j=1}^{N_i} \lambda_{ij}(X_i^b, X_{i,j}^t)^{\delta_{ij}} (1 - \lambda_{ij}(X_i^b, X_{i,j}^t))^{1-\delta_{ij}}. \quad (5.2)$$

As previously mentioned, survival analysis is commonly used in settings where there is right-censoring. [26] proposed inverse-probability-of-censoring-weighting (IPCW) as a method to correct for censoring in cases where the censoring process is (partially) dependent on the baseline and/or time-dependent covariates related to the event process. This approach constructs a pseudo-population for the censored patients by upweighting uncensored patients with similar characteristics. The stabilized weights, W_{ij}^C , are defined as:

$$W_{ij}^C = \begin{cases} \prod_{k=0}^j \frac{\mathbb{P}(C_{ik} = 0)}{\mathbb{P}(C_{ik} = 0 | X_i^b, X_{ik}^t)}, & C_{ij} = 0 \\ 0, & C_{ij} = 1 \end{cases}$$

and can be estimated using a continuation ratio model similar to the outcome model, in which the failure event is censoring. The weights are then incorporated in the outcome model:

$$L = \prod_{i=1}^m \prod_{j=1}^{N_i} W_{ij}^C \lambda_{ij}(X_i^b, X_{ij}^t)^{d_{ij}} (1 - \lambda_{ij}(X_i^b, X_{ij}^t))^{1-d_{ij}}$$

and estimation is performed using weighted logistic regression.

There are two values of interest in this setting that can be estimated using the aforementioned model. First, by taking the exponential of (5.1), the parameter estimates for β and γ can be interpreted as additive effects on the hazard-ratio for experiencing the event in each time interval. Consequently, an $\exp(\beta_p)$ greater (lesser) than 1 (0 in the log form) indicates an upwards (downwards) shift in the hazard for those with higher versus lower values of X_{ip} . The same holds for γ .

The second estimate of interest is restricted mean survival time (RMST). The RMST is the mean survival time restricted at time $t = \tau$:

$$\text{RMST} = \frac{1}{n} \sum_{i=1}^m \prod_{j=1}^{\tau} (1 - \lambda_{ij}(X_i^b, X_{ij}^t)).$$

Hazard ratios, although they indicate whether a covariate is associated with a patient needing more or fewer appointments to reach the desired outcome, do not have a direct interpretation in terms of how many appointments a patient will need. RMST, on the other hand, has a direct interpretation in terms of the number of appointments a patient will

need. Thus, we employ a two-stage approach where we (1) estimate the hazard ratios and (2) use G-computation to estimate the RMST for various levels of the covariates of interest, specifically waiting time and the proportion of appointments delivered in person. The 95% confidence intervals for the association between covariates and RMST are calculated using the Delta method.

5.2 Number of Appointments Needed Cont'd

In order to investigate the relationship between our variables of interest and the number of appointments patients need to achieve desired outcomes, we formulate the functional form of the continuation ratio model as:

$$\text{logit}(\lambda_{ij}) = \alpha_j + \beta_1 \text{Wait}_i + \gamma_1 \text{Face}_{ij} + \gamma_2 \text{Face}_{ij}^2 + \mathbf{Z}_i \delta + \mathbf{V}_{ij} \theta. \quad (5.3)$$

The main effects of interest are those of Wait_i (Hypothesis 1A), Face_{ij} , and Face_{ij}^2 (Hypotheses 2A and 2C). Baseline control variables, including (1) severity, (2) referral age, (3) office location, and (4) the patient's focal clinician are denoted by the vector $\mathbf{Z}_{ij}$. Time-varying controls, including (1) the cumulative average interval between consecutive appointments, (2) the month of the appointment, and (3) the patient's age at appointment j , are included in the vector $\mathbf{V}_{ij}$. Furthermore, we use the same specification as (5.3) for the censoring model that is used to calculate the IPC weights.

To test Hypothesis 2B, whether disorder severity modulates the effect of waiting time and efficiency of in-person versus virtual treatment, we add interaction terms between the severity (Sev_i), waiting time (Wait_i), and the cumulative proportion of treatment that was delivered in-person (Face_{ij} , Face_{ij}^2), resulting in the outcome model formulation:

$$\text{logit}(\lambda_{ij}) = \alpha_j + \beta_1 \text{Sev}_i \times \text{Wait}_i + \gamma_1 \text{Sev}_i \times \text{Face}_{ij} + \gamma_2 \text{Sev}_i \times \text{Face}_{ij}^2 + \mathbf{Z}_i \delta + \mathbf{V}_{ij} \theta.$$

5.3 Cancellation and No-Show Probabilities

Next, we estimate the association between factors such as appointment modality and the number of previously delivered appointments on the probability of a patient no-showing or

canceling their upcoming appointment. In our setting, there are four mutually exclusive outcomes of interest: a patient either attends their appointment, no-shows, cancels at the last minute, or cancels in advance. Furthermore, we have repeated observations for many patients, owing to the multi-appointment nature of care. Thus, we employ a mixed-effects multinomial logistic regression approach to estimate the relationship between the covariates and outcomes. We include random intercepts for each patient to control for unobserved, individual-specific attendance behavior.

To test Hypotheses 3, 4A, and 4B we define the linear predictor of the mean function for outcome j (i.e., $\alpha_j + X\beta_j$) and sample i as l_{ij} :

$$l_{ij} = \alpha_j + \beta_{0j}Modality_i + \beta_{1j}FutureBookings_i + \beta_{2j}FutureBookings_i^2 + \gamma\mathbf{X}_i + \psi\mathbf{Z}_i. \quad (5.4)$$

The main effects of interest included are those of modality, $Modality_i$, and the number of known future bookings, $FutureBookings_i$. In $\mathbf{X}$ we include a variety of scheduling related factors including the linear and quadratic terms of $LeadTime$, Sev , the number of prior appointments, prior cancellation behavior, in addition to the temperature and precipitation on the day of the appointment. Finally, we include a variety of fixed effects in $\mathbf{Z}$, namely: the office location and the weekday, month, and year of the appointment.

To investigate the effect of drive time specifically on in-person appointment attendance, we isolate in-person appointments and add the estimated driving time from the patient's home to their clinic as a covariate to the model defined by (5.4).

Chapter 6

Empirical Results

6.1 Number of Appointments Needed

In Table 6.1 we provide the results for the association between the initial waiting time and modality of treatment on the number of appointments required. Whether or not the dependent censoring mechanism is controlled for using IPCW, the direction of the estimates does not change; however, the magnitude of the association shifts marginally after dependent censoring is accounted for. Overall, waiting times are positively associated with patients needing more appointments, as evidenced by the 57.2% reduction in the odds of discharge per year of waiting time. This observation supports Hypothesis 1A. Furthermore, using a two-stage estimation method, we estimate the average marginal effect (AME) of waiting time on RMST (restricted at 15 appointments) is equal to patients needing 2.94 (26.3%) more appointments for each year of waiting time. In Figure 6.1(a), we present the graphical results of this estimation, including the 95% confidence interval calculated using the Delta method.

Our analysis also shows a significant, quadratic relationship between modality and the number of appointments patients require. However, examining the RMST estimation we see that this is not necessarily an inverted U-shaped relationship, where mixing modalities (e.g., 50% in-person treatment) results in patients needing significantly more appointments than uni-modular treatment. Rather, there appears to be little variance in the efficiency of treatment when 50% or less treatment is in-person. Comparatively, efficiency significantly increases as more treatment, beyond the 50% threshold, is given in-person. Specifically, receiving all treatment in-person is associated with patients needing 0.94 (8.0%) fewer appointments compared to all virtual treatment.

Table 6.1: Survival Analysis Model Results

Variable	OR	OR (IPCW)	OR	OR (IPCW)
Sev_L	0.025 (0.052)	0.043 (0.093)	0.028 (0.057)	0.046 (0.097)
Sev_M	0.008 (0.018)	0.014 (0.030)	0.009 (0.019)	0.015 (0.032)
Sev_H	0.003 (0.007)	0.006 (0.012)	0.004 (0.008)	0.006 (0.013)
$WaitTime$	0.420*** (0.118)	0.428*** (0.131)	0.392*** (0.109)	0.411*** (0.126)
$PctFace$	1.662*** (0.299)	1.563*** (0.280)	0.524* (0.294)	0.777* (0.446)
$PctFace^2$			3.468* (1.980)	2.130* (1.243)
Controls	Included	Included	Included	Included
Observations	15806	15488	15806	15488

Notes. The standard errors are reported in parentheses, with *, **, and *** indicating significance at the 0.05, 0.01, and 0.001 levels, respectively.

As previously explained, we have reason to believe the effect of wait time and modality may differ between severity levels. To investigate this, we add interaction terms between wait time ($WaitTime$), modality ($PctFace$), and severity (Sev). The results for the models with and without IPCW, as well as with and without the quadratic effect of modality, are provided in Table 6.2. Considering the model including only the linear terms of modality, we find that modality is only significantly associated with the number of appointments patients receive if they have a medium or high severity disorder. These patients need fewer appointments to achieve the desired outcomes if they receive more of their treatment in-person. From Figure 6.2(b), we see that there is no significant association between modality and the number of appointments required for low severity patients. However, when we consider the quadratic term of modality, we see similar results for the medium

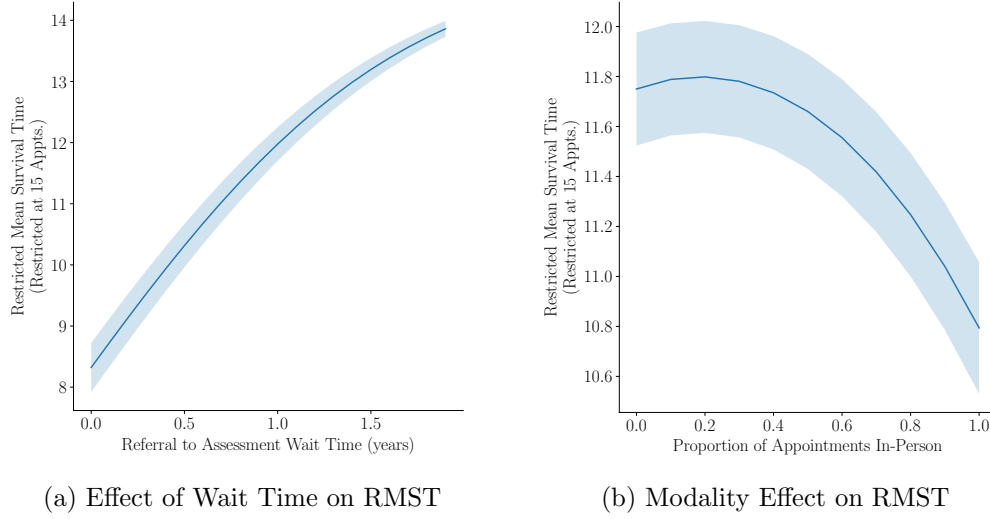


Figure 6.1: Estimation of the Effect of Wait Time and Modality on Restricted Mean Survival Time (RMST)

and high severity patients as in the non-stratified case. That is to say, in-person treatment is significantly more efficient than virtual treatment, but only when the majority of visits are delivered in-person. In the case of medium-severity patients, receiving all appointments in-person is associated with a 19.35% reduction in the number of appointments needed compared to fully virtual treatment. Similarly, high-severity patients benefit from an 18.04% service efficiency improvement when all treatment is in-person compared to virtual.

When considering the effect of wait time stratified by disorder severity, we find that medium and high severity patients require more appointments the longer they wait. This is supported by 1 year of waiting time decreasing the odds of being discharged by 69.3% and 60.2%, respectively, for medium and high severity patients. In terms of RMST, a 1 year increase in waiting time is associated with medium and high severity patients needing, respectively, 2.4 (21%) and 1.2 (9%) more appointments. On the other hand, wait time is observed to have no significant effect on the number of appointments low severity patients need.

Table 6.2: Survival Analysis Model Results Stratified by Severity Level

Variable	OR	OR (IPCW)	OR	OR (IPCW)
$WaitTime \times Sev_L$	0.611 (0.207)	0.633 (0.229)	0.651 (0.222)	0.694 (0.253)
$WaitTime \times Sev_M$	0.307*** (0.095)	0.297*** (0.099)	0.288*** (0.089)	0.284*** (0.094)
$WaitTime \times Sev_H$	0.398** (0.160)	0.375** (0.171)	0.336*** (0.138)	0.318** (0.145)
$PctFace \times Sev_L$	0.849 (0.218)	0.873 (0.222)	5.674* (5.133)	8.154** (7.443)
$PctFace \times Sev_M$	2.324*** (0.571)	2.096*** (0.507)	0.461 (0.359)	0.589 (0.465)
$PctFace \times Sev_H$	3.429** (1.867)	2.851** (1.516)	0.019*** (0.026)	0.028** (0.040)
$PctFace^2 \times Sev_L$	-	-	0.098** (0.103)	0.066** (0.070)
$PctFace^2 \times Sev_M$	-	-	5.248** (4.076)	3.736* (2.939)
$PctFace^2 \times Sev_H$	-	-	196.143*** (255.317)	120.008*** (168.522)
Controls	Included	Included	Included	Included
Observations	15,806	15,488	15,806	15,488

Notes. The standard errors are reported in parentheses, with *, **, and *** indicating significance at the 0.05, 0.01, and 0.001 levels, respectively.

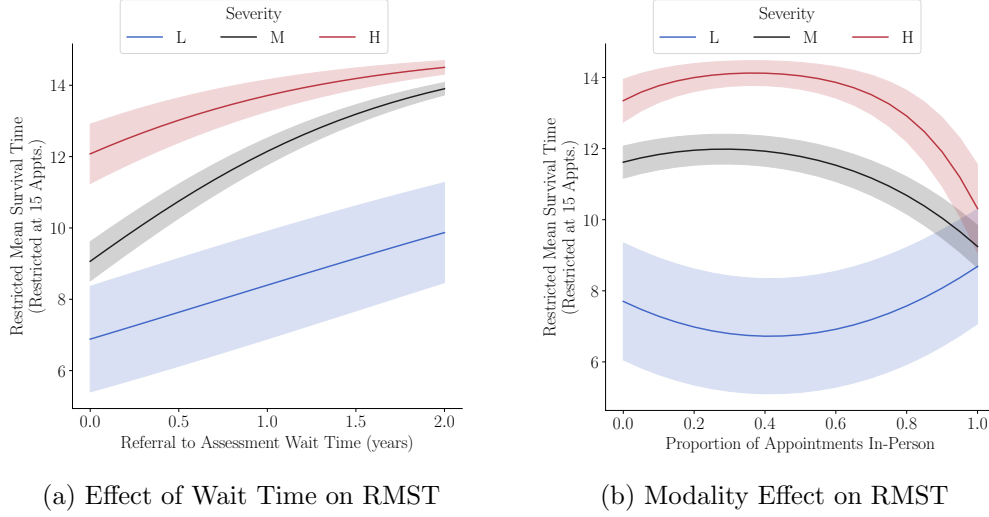


Figure 6.2: Estimation of the Effect of Wait Time and Modality on Restricted Mean Survival Time (RMST), Stratified by Severity Level

6.2 Cancellation and No-Show Probabilities

Next, we will consider factors that are associated with the likelihood of attendance, no-show, last-minute cancellation, or advance cancellation of an appointment. We note that of the various types of non-attendance, no-shows and last-minute cancellations have the greatest negative effect on effective treatment capacity usage, with these vacant appointment slots going unfilled by another patient 98.3% and 97.8% of the time in our data. Unlike other healthcare settings, such as family medicine, there are unique constraints on what approaches can be used to mitigate the effect of no-shows and last-minute cancellations in SLP. As patients must be treated by their primary clinician and contract constraints preclude overtime work, overbooking is extremely difficult to implement effectively in this setting. For this reason, limiting the rate at which no-shows and last-minute cancellations occur is more effective in practice than mitigating their negative effects on throughput after they occur.

We begin by estimating the effects of modality, the number of appointments already received by a patient (denoted by *PrevAppt*), and the number of future bookings known to the patient at the time of the focal attendance event (denoted by *FutureAppts*). The results for the multinomial logistic regression model including these effects and fixed effects for the control variables are presented in Table 6.3. Beginning with the effect of modality

on attendance outcome likelihoods, we find that virtual appointments are statistically significantly associated with a lower likelihood of a no-show ($OR = 0.575$, $p < 0.001$) or last-minute cancellation ($OR = 0.661$, $p < 0.001$) occurring, compared to in-person appointments. Examining the AME of modality on the probability of each outcome, we observe that virtual appointments are 4.42% more likely to be attended than in-person appointments. This is driven by significantly lower probabilities of a no-show or last-minute cancellation occurring. Virtual appointments have an average marginal effect on reducing the probability of no-shows and last-minute cancellations of 1.88% and 3.17%, respectively. However, we do not observe a significant association between modality and the likelihood of an advance cancellation.

Table 6.3: Multinomial Logistic Model Results

	<i>Attendance outcome:</i>		
	No-Show	Cancellation (≤ 24 Hours)	Cancellation (> 24 Hours)
	(1)	(2)	(3)
<i>Modality</i> =Virtual	0.575*** (0.122)	0.661*** (0.096)	0.866 (0.112)
<i>FutureBookings</i>	0.833* (0.085)	1.202** (0.064)	5.300*** (0.080)
<i>FutureBookings</i> ²	1.029 (0.024)	0.965 (0.017)	0.720*** (0.021)
Controls	Included	Included	Included
Observations	22,169	22,169	22,169

Notes. The $\log(OR)$ standard errors are reported in parentheses, with *, **, and *** indicating significance at the 0.05, 0.01, and 0.001 levels, respectively. Primary values presented in each column are the odds ratio (OR) estimates from the multinomial logistic regression model for the respective non-attendance type outcome compared to attendance.

Turning our attention to the effect of the number of known future bookings on the probability of attendance outcomes, we observe a statistically significant linear relationship for all types of non-attendance. However, there is only a significant quadratic relationship with

advanced cancellation. Although these are statistically significant relationships, their economic significance is questionable for some attendance outcomes. In the case of no-shows, an extra future known booking decreases the probability of a no-show by 0.18% in absolute terms. Relative to the 4.85% average rate of no-showing, this is a 3.7% relative reduction. However, an extra future known booking negligibly decreases the likelihood of a last-minute cancellation. Future known bookings do, however, exhibit an important, inverted-U-shaped relationship with advanced cancellation behavior. Between 0 and 3 future bookings, the likelihood of an advanced cancellation increases by 10%, however, appointments with 5 future known bookings are associated with only a 0.4% greater likelihood of resulting in an advanced cancellation.

Although having a greater number of appointments booked in the future is associated with a decreased likelihood of the upcoming appointment resulting in a no-show, it is also associated with a greater likelihood of non-attendance for future bookings. As seen in Figure A.1, lead times for future known bookings increase monotonically. Appointment lead time has been established to have a positive association with appointment non-attendance. Due to this relationship and the positive association between lead time and non-attendance behavior, although booking more appointments into the future decreases the likelihood of the upcoming appointment resulting in a no-show, the likelihood of these future appointments being attended also decreases due to the associated increase in their lead times.

The last primary effect that we are interested in is the effect of the estimated driving time from a client’s home to their clinic on attendance outcome likelihoods. We use the same model as before; however, we add the estimated driving time, measured in minutes, as a regressor and estimate the parameters using only the in-person appointment bookings. The results for this model are presented in Table 6.4. We make note that the relationship for the number of known future bookings does not change significantly from the model including both in-person and virtual appointments. However, the minimal association with reduced no-show likelihood is no longer statistically significant. This suggests that these effects are largely not modulated by modality.

Focusing on the effect of drive time, we find that a longer estimated drive time between a patient’s home and the clinic is associated with a decreased likelihood of a no-show occurring but an increased likelihood of an advance cancellation occurring. Examining the AMEs, we find that each additional minute of driving is associated with a 0.14% decrease ($p < 0.001$) in the probability of a no-show and a 0.09% increase ($p < 0.01$) in the likelihood of an advanced cancellation taking place. Although the overall probability of attendance is not significantly associated with driving time, these results indicate that patients who live further away from their clinic are more likely to give advance notice that they will not attend their appointment compared to those living within a short driving distance.

Table 6.4: Multinomial Logistic Model Results - In-Person Only

	<i>Attendance outcome</i>		
	No-Show	Cancellation (≤ 24 Hours)	Cancellation (> 24 Hours)
	(1)	(2)	(3)
<i>DriveTime</i>	0.971*** (0.007)	0.998 (0.005)	1.020*** (0.007)
<i>FutureBookings</i>	0.843 (0.093)	1.201*** (0.071)	5.232*** (0.091)
<i>FutureBookings</i> ²	1.023 (0.029)	0.968 (0.021)	0.722*** (0.025)
Controls	Included	Included	Included
Observations	19,365	19,365	19,365
Pseudo R^2	0.186	0.186	0.186

Notes. The standard errors are reported in parentheses, with *, **, and *** indicating significance at the 0.05, 0.01, and 0.001 levels, respectively. Primary values presented in each column are the odds ratio (OR) estimates from the multinomial logistic regression model for the respective non-attendance type outcome compared to attendance.

Chapter 7

Counterfactual Analysis

In Chapter 7, we use a queueing model of patient flow to evaluate the effects of various operational interventions on system throughput and patient care. We focus on long-run average waiting times, appointment fill rates, and age-out rates as the performance measures. In Section 7.1, we describe the simulation design. In Section 7.2, we focus specifically on interventions that account for the effects of wait time on the number of appointments that patients need. In Section 7.3, we examine policies that account for the effects of modality on the number of appointments patients need and their attendance behavior.

7.1 Simulation Design

The simulation is set up as a queueing system, with multiple parallel servers with individual queues that each represent a weekly appointment slot. The size of each server's queue determines the appointment frequency of the patients on their panel (i.e., a rotating queue of four patients equates to monthly service for each patient). Servers deliver an appointment to the patient at the front of their respective queue each week and subsequently either: 1) discharge this patient if they have received all their needed appointments or aged out of the system, or 2) place them at the back of their queue. The waitlist is represented either by a) a single first-in-first-out (FIFO) queue in the case of FCFS service, or b) separate FIFO queues for each patient class, used when simulating strict priority policies for assigning patients to servers. When a server discharges a patient, they pull a new patient from the waitlist according to the waitlist policy. Furthermore, we assume that the servers are versatile and can accommodate patients of any class.

For the simulations presented in the subsequent sections, we focus on data-driven parameterization. We assume that the weekly number of arrivals is Poisson-distributed, with a mean of 32. The number of weekly appointment slots is 261 and is constant over time. In addition, patients receive treatment with a monthly frequency. The number of appointments patients need and the effects of wait time and modality are calibrated using the RMST estimations in Section 6.1.

7.2 Effects of Wait Time

Determining the order in which to add patients from the waitlist is an important strategic decision in the control of multiclass many-server queues. Our queueing model can help optimize these decisions. We use our model to estimate the differences in performance between a single FCFS waitlist and strict-priority policies in a setting where service rates are dependent on waiting times and are disparate across classes.

Table 7.1: Arrival Rates, Baseline Service Rates, and Wait Time Effects for Three Severity Classes

Severity Class	Arrival Rate	Baseline Appointments Needed	Extra Appointments (per year waited)	Extra Appointments (100% In-Person)
L	4.99	7	0	0
M	15.87	9	2.4	-1.436
H	9.14	12	1.2	-1.115

Based on the estimation results in Section 6.1, we parameterize the model according to Table 7.2. Ignoring the effect of waiting time, we note that a fully data-driven parameterization results in an approximate utilization (ρ) of 1.09. We examine this case and lower levels of utilization. It is important to note that, because of the aging-out process, the system behaves somewhat differently when over-utilized (i.e., $\rho > 1$) compared to other queueing systems.

Figure 7.1 shows how the optimal policy, in terms of the minimal total cost, changes as the cost of aging-out/waiting time changes between the three severity classes. It also shows the relative differences between the strict priority and FCFS policies. Overall, the strict priority policy significantly outperforms the FCFS policy in all cost settings, resulting in a relative reduction of total cost between 50% and 95%. In Figure 7.1(a) we see that, when the classes' costs are equal, the best strict priority policy is to accept patients from

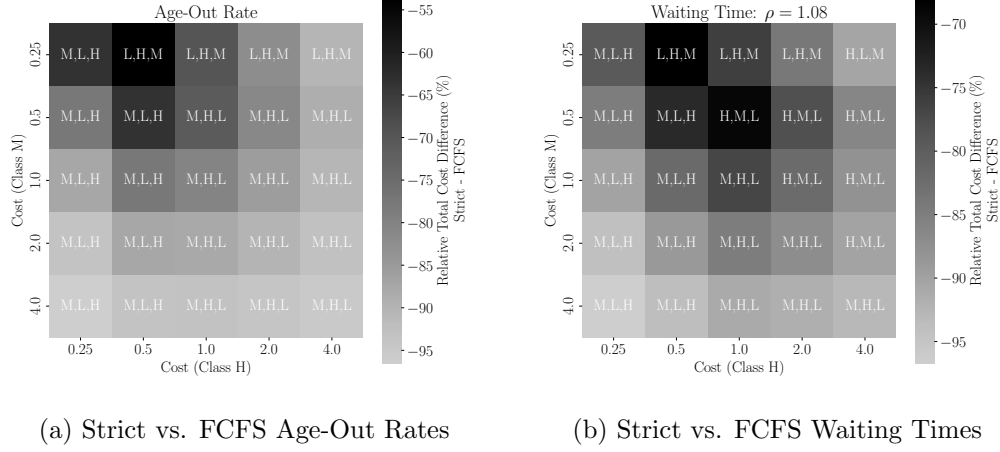


Figure 7.1: Best Performing Strict Priority Policies Based on Class Costs for Long-Run Age-Out Rates and Waiting Times (Class L Cost Normalized to 1)

the waitlist in order of medium, high, and then low severity. This policy results in an 80% reduction in the age-out rate compared to the FCFS policy. We note that there is no significant difference in the age-out costs between this policy and a high, medium, then low ordering. This policy remains the best choice if considering waiting times instead of age-out rates, resulting in a 76.2% reduction in waiting times. These observations contradict the $c\mu$ rule, which would propose a low, medium, then high ordering. Furthermore, a high, medium, and then low severity priority order further significantly reduces age-out rates and waiting times by 45.87% and 4.71%, respectively, compared to a low, medium, then high severity order. Comparatively, when there is no effect of waiting time on the number of appointments patients need, the $c\mu$ rule holds, both when minimizing the cost of aging-out and waiting.

7.3 Effects of Modality

As was elucidated in Section 6, the modality of treatment is associated with both the number of appointments patients require and the likelihood of them attending a given appointment. Furthermore, the relationship between the number of appointments needed and appointment modality is modulated by patient severity. Currently, low, medium, and high severity patients receive approximately 82%, 87%, and 87% of their appointments in-person, respectively.

Table 7.2: Simulation Results for Different Modality Policies

Modality Policy	FCFS (No Wait Effect)		M,H,L Priority Order	
	Age-Out Rate	Waiting Time	Age-Out Rate	Waiting Time
0.0,0.0,0.0	0.67 ± 0.00	77.45 ± 0.05	0.46 ± 0.00	25.20 ± 0.04
0.8,0.9,0.9	0.46 ± 0.00	75.95 ± 0.05	$0.18^\dagger \pm 0.01$	$7.11^\dagger \pm 1.35$
0.0,1.0,1.0	$0.41^\dagger \pm 0.00$	$75.07^\dagger \pm 0.08$	$0.17^\dagger \pm 0.00$	$9.58^\dagger \pm 0.45$
1.0,1.0,1.0	0.43 ± 0.00	75.44 ± 0.07	$0.17^\dagger \pm 0.00$	$9.19^\dagger \pm 0.32$

Notes. Results are the averages across all three patient classes. Margins of error at a 95% confidence level are reported in parentheses.

Table 7.2 presents the results for different proportions of in-person treatment given to each class under two different parametrization settings: 1) when the waitlist is FCFS and wait time has no effect on the number of appointments needed, and 2) when the waitlist follows a medium, high, then low (M,H,L) priority ordering and wait time has an effect. The M,H,L priority ordering is displayed as it resulted in the lowest age-out rates and waiting times of the possible priority orderings. In the FCFS setting, we see that the optimal modality policy, which is to provide 100% virtual treatment to low severity patients, and 100% in-person treatment to medium and high severity patients, significantly reduces the proportion of patients that age-out by 0.05 (10.9%) compared to the data-driven modality policy.

In the setting where wait times affect the number of appointments needed, and the waitlist follows a strict priority ordering (M,H,L), we do not see a significant difference between the best modality policy (0,1,1) and the data-driven policy (0.8,0.9,0.9). We argue this is because the change induced by the priority ordering overwhelms the effects of the modality policy, resulting in marginal differences. There is still, however, a clear benefit to intelligently picking a modality policy. For example, naively providing all treatment virtually would result in a 170% and 163% increases, respectively, in the average age-out rate and waiting time compared to the optimal policy.

At a realistic utilization ($\rho = 1.09$), differences between modality policies are marginal when the wait times affect the number of appointments required and the waitlist operates in a FCFS fashion. However, when the system operates at a slightly underutilized level ($\rho = 0.975$), modality policy is observed to significantly affect age-out rates. In the context of slight under-utilization, the optimal policy is 100% virtual treatment for low severity patients and 100% in-person treatment for medium and high severity patients; this trans-

lates to a 0.076 (9.6%) reduction in the proportion of patients that age-out compared to the data-driven modality policy.

Chapter 8

Conclusion

This paper investigates the effects of operational factors and scheduling decisions on system throughput in a multi-appointment, outpatient care setting. Specifically, we investigate the impacts of initial waiting time and treatment modality on the number of treatments patients need to achieve desired outcomes, as well as the effect of appointment modality on appointment attendance behavior. Using appointment-level scheduling data, we find that patients require more treatments when they wait longer to begin treatment or receive more of their treatment via telehealth methods. However, these effects are modulated by patient severity. Our findings highlight the importance of considering how scheduling decisions can affect a patient’s treatment progression, behavior, and overall system throughput. Although service delay impacts on service duration have been investigated before, that research is primarily in single-appointment care settings where delays worsen patient condition (e.g., [6]). Similarly, operations management research pertaining to the efficacy of telehealth has focused on single-appointment care settings, where telehealth usage can increase the necessity for a follow-up visit. Our setting differs in that we examine telehealth usage in the context of multi-appointment rehabilitation care and the total number of treatments patients require.

We begin by investigating the effect of initial waiting time and telehealth usage on the number of treatments patients need to achieve desired outcomes. We find that, for medium- and high-severity patients, one year of waiting is associated with a 21% and 9% increase, respectively, in the number of treatments required compared to no waiting time. However, for low-severity patients we observe no significant effect of waiting time on the number of treatments needed. We posit that in the setting of SLP, and rehabilitation services more broadly, longer waiting times and greater telehealth usage may slow medium- and high-severity patients’ progress, leading to more treatments being required to achieve the

desired outcome. A similar pattern emerges when we consider the relationship between telehealth usage and the number of appointments needed. Whereas medium- and high-severity patients benefit significantly from in-person treatment (that is, they need fewer total treatments when a greater proportion of treatments are delivered in-person), low-severity patients do not appear to be significantly affected by the modality of treatment.

Although in-person treatments can be more effective from a treatment perspective for some patient categories, our appointment attendance behavior analysis shows that in-person treatments are also 4.42% less likely to be attended by patients. This emphasizes the importance of considering the effects of scheduling and treatment decisions on multiple drivers of system throughput, rather than myopically focusing on the number of appointments patients need or their attendance behavior. In addition to studying the relationship between telehealth and attendance behavior, we also investigate the association between the number of known future bookings a patient has and their likelihood of keeping their upcoming appointment. We find that the more appointments a patient has booked in the future, the lower their likelihood of no-showing their upcoming appointment becomes. This may be attributable to a no-show deterrent policy where patients are removed from service if they no-show more than two appointments. Thus, if a patient has several future bookings already made, there is a known consequence to no-showing their next appointment. Secondly, we find that patients are significantly more likely to cancel their next appointment in advance (>24 hours) up to 3 future bookings. One explanation for this behavior is that patients are not directly penalized by the organization for canceling appointments with advance notice. Therefore, when a patient has more known future bookings they are less averse to canceling an upcoming appointment as they have a known amount of guaranteed service remaining.

Lastly, we examine the managerial implications of the effects of waiting time and telehealth on throughput. Using counterfactual simulations, we find that using a strict priority policy waiting times can be decreased by 76.2% compared to the currently used FCFS policy. Furthermore, the best strict priority policy does not align with the $c\mu$ -rule in the presence of waiting time effects on the number of appointments needed. Secondly, we show that when telehealth affects both the number of appointments patients need as well as their appointment attendance behavior, the choice of how frequently different patient classes should be given virtual appointments depends on the magnitude of both of these effects. Importantly, we show that compared to the historical virtual appointment allocation policy, age-out rates can be decreased relatively by 10.9% through improved decision-making regarding the allocation of virtual treatment.

There are several opportunities for future research. First, our study is limited by the use of retrospective data. Although we generate hypotheses to explain why waiting

time and virtual appointments affect the number of appointments patients need in a multi-appointment healthcare setting, future work is necessary to validate these hypotheses. The same holds for our study of appointment attendance behavior. Given the unique nature of pediatric rehabilitation care, we invite other researchers to study the effects of waiting time and telehealth usage in other multi-appointment healthcare settings. We hope that by showing scheduling decisions can differentially impact multiple drivers of throughput (e.g., whether to deliver appointments virtually or in-person affects both the number of appointments needed and attendance behavior), researchers will develop scheduling models that consider these effects jointly, as opposed to myopically focusing on individual drivers of throughput such as service duration or no-show likelihood.

References

- [1] Hessam Bavafa, Lorin M. Hitt, and Christian Terwiesch. The Impact of E-Visits on Visit Frequencies and Patient Health: Evidence from Primary Care. *Management Science*, 64(12):5461–5480, December 2018.
- [2] Andrew B. Bean and James Talaga. Predicting appointment breaking. *Journal of Health Care Marketing*, 15(1):29, 1995.
- [3] Dependra Bhatta, Monteic A. Sizer, and Binod Acharya. Association Between Telehealth and Missed Appointments Among Patients Experiencing Behavioral Health Challenges. *JAMA Network Open*, 6(7):e2324252, July 2023.
- [4] Nina Capone Singleton. Late Talkers: Why the Wait-and-See Approach Is Outdated. *Pediatric Clinics of North America*, 65(1):13–29, February 2018.
- [5] Donald B. Chalfin, Stephen Trzeciak, Antonios Likourezos, Brigitte M. Baumann, R. Phillip Dellinger, and for the DELAY-ED study Group. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit*. *Critical Care Medicine*, 35(6):1477, June 2007.
- [6] Carri W. Chan, Vivek F. Farias, and Gabriel J. Escobar. The Impact of Delays on Service Times in the Intensive Care Unit. *Management Science*, 63(7):2049–2072, July 2017.
- [7] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.
- [8] Richard A. Deyo and Thomas S. Inui. Dropouts and Broken Appointments: A Literature Review and Agenda for Future Research. *Medical Care*, 18(11):1146, November 1980.

- [9] Jing Dong and Rouba Ibrahim. SRPT Scheduling Discipline in Many-Server Queues with Impatient Customers. *Management Science*, 67(12):7708–7718, December 2021.
- [10] Marge Edwards, Arlene Stredler-Brown, and K. Todd Houston. Expanding Use of Telepractice in Speech-Language Pathology and Audiology. | Volta Review | EBSCOhost, December 2012. ISSN: 0042-8639 Issue: 3 Pages: 227 Volume: 112.
- [11] Elahe Farmani, Fatemeh Fekar Gharamaleki, and Mohammad Amin Nazari. Challenges and opportunities of tele-speech therapy: Before and during the COVID-19 pandemic. *Journal of Public Health Research*, 13(1):22799036231222115, January 2024.
- [12] Aliza S. Gordon, Wallace C. Adamson, and Andrea R. DeVries. Virtual Visits for Acute, Nonurgent Care: A Claims Analysis of Episode-Level Utilization. *Journal of Medical Internet Research*, 19(2):e6783, February 2017.
- [13] Michael Gordon, Kevin M. Antshel, Lawrence Lewandowski, and Danielle Seigers. Economic Grand Rounds: Predictors of Missed Appointments Over the Course of Child Mental Health Treatment. *PS*, 61(7):657–659, July 2010.
- [14] Elizabeth Hesketh, Paul White, Doug Simkiss, and Sue Roulstone. Outcomes from a community speech and language therapy service treatment waiting list: The natural history of 525 children with identified speech and language needs. *International Journal of Language & Communication Disorders*, 58(5):1510–1525, 2023.
- [15] Suparerk Lekwijit, Christian Terwiesch, David A. Asch, and Kevin G. Volpp. Evaluating the Efficacy of Connected Healthcare: An Empirical Examination of Patient Engagement Approaches and Their Impact on Readmission. *Management Science*, 70(6):3417–3446, June 2024.
- [16] Nan Liu. Optimal Choice for Appointment Scheduling Window under Patient No-Show Behavior. *Production and Operations Management*, 25(1):128–142, January 2016.
- [17] Nan Liu, Shan Wang, and Noa Zychlinski. RL or URL: Managing Outpatient (Tele)visits with Strategic Behavior, March 2023.
- [18] Stephen Martin and Peter C. Smith. Rationing by waiting lists: an empirical investigation. *Journal of Public Economics*, 71(1):141–164, January 1999.
- [19] Emma McLaughlin, Michelle Lincoln, and Barbara Adamson. Speech-language pathologists’ views on attrition from the profession. *International Journal of Speech-Language Pathology*, 10(3):156–168, June 2008.

- [20] Anya J. Miller, Esther Chae, Ed Peterson, and Alvin B. Ko. Predictors of repeated “no-showing” to clinic appointments. *American Journal of Otolaryngology*, 36(3):411–414, May 2015.
- [21] Khadijeh Moulaei, Abbas Sheikhtaheri, Farhad Fatehi, Mostafa Shanbehzadeh, and Kambiz Bahaadinbeigy. Patients’ perspectives and preferences toward telemedicine versus in-person visits: a mixed-methods study on 1226 patients. *BMC Medical Informatics and Decision Making*, 23(1):261, November 2023.
- [22] John B. Norris, Chetan Kumar, Suresh Chand, Herbert Moskowitz, Steve A. Shade, and Deanna R. Willis. An empirical investigation into factors affecting patient cancellations and no-shows at outpatient clinics. *Decision Support Systems*, 57:428–443, January 2014.
- [23] Chinedum O. Ojinnaka, Lara Johnstun, Anthony Dunnigan, Lora Nordstrom, and Sandra Yuh. Telemedicine Reduces Missed Appointments but Disparities Persist. *Am J Prev Med*, 67(1):90–96, July 2024.
- [24] Balaraman Rajan, Tolga Tezcan, and Abraham Seidmann. Service Systems with Heterogeneous Customers: Investigating the Effect of Telemedicine on Chronic Care. *Management Science*, 65(3):1236–1267, March 2019.
- [25] Benjamin Ravenscroft, Hossein Abouee Mehrizi, and Brendan Wylie-Toal. Effects of Booking Horizon Reduction on Cancellation Rates: An Experimental Analysis in Pediatric Outpatient Care. *MDM Policy & Practice*, 9(2):23814683241298673, July 2024.
- [26] James M. Robins and Dianne M. Finkelstein. Correcting for Noncompliance and Dependent Censoring in an AIDS Clinical Trial with Inverse Probability of Censoring Weighted (IPCW) Log-Rank Tests. *Biometrics*, 56(3):779–788, 2000.
- [27] P. B. Seetharaman. The Additive Risk Model for Purchase Timing. *Marketing Science*, 23(2):234–242, May 2004.
- [28] Yunting Shi, Nan Liu, and Guohua Wan. Treatment Planning for Victims with Heterogeneous Time Sensitivities in Mass Casualty Incidents. *Operations Research*, 72(4):1400–1420, July 2024.
- [29] Zhuo Shi, Ateev Mehrotra, Courtney A. Gidengil, Sabrina J. Poon, Lori Uscher-Pines, and Kristin N. Ray. Quality Of Care For Acute Respiratory Infections During

- Direct-To-Consumer Telemedicine Visits For Adults. *Health Affairs*, 37(12):2014–2023, December 2018.
- [30] Katie Squires. Addressing the Shortage of Speech-Language Pathologists in School Settings. *Journal of the American Academy of Special Education Professionals*, 2013.
 - [31] Maria Stepanova and Lyn Thomas. Survival Analysis Methods for Personal Loan Data. *Operations Research*, 50(2):277–289, April 2002.
 - [32] Adam S. Tenforde, Haylee Borgstrom, Ginger Polich, Hannah Steere, Irene S. Davis, Kester Cotton, Mary O’Donnell, and Julie K. Silver. Outpatient Physical, Occupational, and Speech Therapy Synchronous Telemedicine: A Survey Study of Patient Satisfaction with Virtual Visits During the COVID-19 Pandemic. *Am J Phys Med Rehabil*, page 10.1097/PHM.0000000000001571, August 2020.
 - [33] Danielle Wales, Leisa Skinner, and Melanie Hayman. The Efficacy of Telehealth-Delivered Speech and Language Intervention for Primary School-Age Children: A Systematic Review. *Int J Telerehabil*, 9(1):55–70, June 2017.
 - [34] Özden Engin Çakıcı and Alex F. Mills. Telehealth in Acute Care: Pay Parity and Patient Access. *M&SOM*, November 2024.

APPENDICES

Appendix A

Additional Figures and Tables

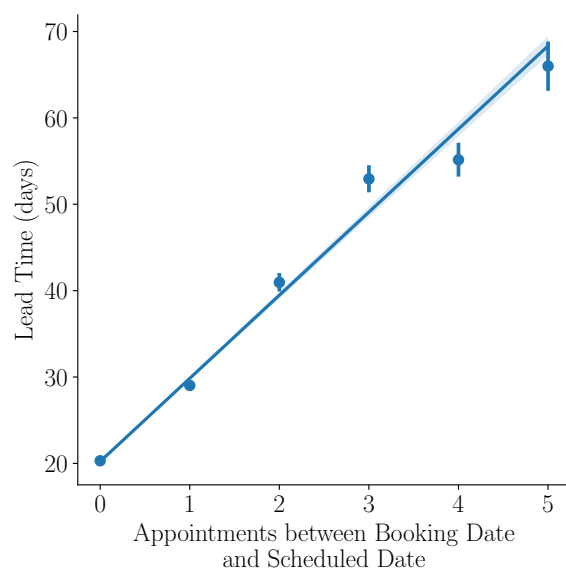


Figure A.1: Relationship Between Lead Time and the Number of Appointments Scheduled Between When the Focal Appointment was Booked and When it Was Scheduled to Occur.

Table A.1: Standardized Mean Difference of Covariates Between In-Person and Virtual Appointments

Variable	SMD
Future Known Bookings	0.09
Lead Time (days)	0.07
Drive Time (min)	0.02
Previous Attendances	0.19
Previous Attendances ²	0.14
Previous Cancellations (< 24 hours)	0.13
Previous Cancellations ($\geq$ 24 hours)	0.04
Temperature	0.08
Rainfall (mm)	-0.04
Snowfall (mm)	-0.12

Table A.2: Multinomial Logistic Model Results - Average Marginal Effects

	<i>Attendance outcome:</i>		
	No-Show	Cancellation ($\leq$ 24 Hours)	Cancellation ($>$ 24 Hours)
	(1)	(2)	(3)
<i>Modality</i> =Virtual	-0.0188***	-0.0317***	0.0063
Controls	Included	Included	Included
Observations	22,169	22,169	22,169

Table A.3: Multinomial Logistic Model Results - Average Marginal Effects (In-Person Appointments)

	<i>Attendance outcome:</i>		
	No-Show	Cancellation (≤ 24 Hours)	Cancellation (> 24 Hours)
	(1)	(2)	(3)
<i>DriveTime</i>	-0.0014***	-0.0001	0.0009**
Controls	Included	Included	Included
Observations	22,169	22,169	22,169

Table A.4: Simulation Results for Different Modality Policies - $\rho = 0.97$

Modality Policy	Age-Out Rate	Waiting Time
0.0,0.0,0.0	0.39 ± 0.005	70.67 ± 0.129
0.8,0.8,0.8	0.13 ± 0.006	61.26 ± 0.703
0.0,1.0,1.0	$0.03^{\dagger} \pm 0.012$	$45.87^{\dagger} \pm 4.080$
1.0,1.0,1.0	0.07 ± 0.014	54.63 ± 3.730

Notes. Averages and 95% confidence intervals are reported. Modality policies are reported for patient classes in order of severity. $\dagger$ indicates the best performing policy and policies that are not statistically significantly different from the best performing policy (if applicable).