Waiting Online versus In-Person: An Empirical Study on Outpatient Clinic Visit Incompletion

Jimmy Qin, Carri W. Chan, Jing Dong
Decision, Risk, and Operations, Columbia Business School
qqin23@gsb.columbia.edu, cwchan@gsb.columbia.edu, jing.dong@gsb.columbia.edu

Shunichi Homma, Siqin Ye
Department of Medicine, Columbia University Irving Medical Center
sh23@cumc.columbia.edu, sy2357@cumc.columbia.edu

The adoption of online services, such as telemedicine, has increased rapidly over the last few years. To better manage online services and effectively integrate them with in-person services, we need to better understand customer behaviors under the two service modalities. Utilizing data from two large internal medicine outpatient clinics, we take an empirical approach to study service incompletion for in-person and telemedicine appointments respectively. We focus on estimating the causal effect of physician availability on service incompletion. When physicians are unavailable, patients may be more likely to leave without being seen. We introduce a multivariate probit model with instrumental variables to handle estimation challenges due to endogeneity, sample selection, and measurement error. Our estimation results show that intra-day delay increases the telemedicine service incompletion rate by 7.40%, but it does not have a significant effect on the in-person service incompletion rate. This suggests that telemedicine patients may leave without being seen when delayed, while in-person patients are not sensitive to intra-day delay. We conduct counterfactual experiments to optimize the intra-day sequencing rule when having both telemedicine and in-person patients. Our analysis indicates that not correctly differentiating the types of incompletions due to intra-day delays from no-show can lead to highly suboptimal patient sequencing decisions.

Key words: empirical study; online services; telemedicine; no-show; abandonment

1. Introduction

The adoption of online services is becoming ubiquitous. Spurred by the COVID-19 pandemic, online service now often complements, and in some instances has replaced, in-person service interactions even in industries – e.g., education (World Economic Forum and Wood 2022), law (Statista 2021), financial advising (McKinsey & Company 2015), and healthcare (Patel et al. 2021) – which have traditionally relied predominantly (or even exclusively) on in-person engagements. While there are certainly many benefits that come with the online modality, there are still many open questions on the operationalization of integrating online and in-person modalities. Critical to this is a need to understand whether and how customer behaviors vary across service modalities. In this work, we use in-person and

telehealth outpatient visits as a canonical setting to study how the service completion behavior of customers (patients) depends on visit modality and the implications of this on how to integrate in-person and online (telehealth) appointments.

Telemedicine is one area that has exhibited extensive growth during the pandemic, with pre-pandemic levels of telehealth visits constituting less than 0.1% of all outpatient visits contrasting with up to 25% of Americans using telemedicine to meet their medical needs during the peak of the pandemic (Karimi et al. 2022). There are a number of advantages of virtual visits such as zero commuting time for patients and reduced waiting room density (Wootton et al. 2011, Russo et al. 2016, Paquette and Lin 2019). Thus, many practitioners expect that telemedicine will continue to be integrated with traditional in-person visits to become a routine medical resource delivery modality (Friedman et al. 2022).

How to utilize and integrate virtual visits with in-person visits remains a challenge (Bayram et al. 2020, Lokken et al. 2020). In practice, these decisions have generally been made based on clinical and administrative judgment, with limited use of evidence-based approaches. Yet, such decisions likely have a great impact on access and service quality. When designing an appointment system, many factors are taken into consideration such as customer punctuality and service duration. In this work, we focus on service incompletion, which refers to the phenomenon when the customer does not attend a scheduled appointment. There are substantial consequences of service incompletion, including reduced revenue, provider idle time, and potentially worse service quality. In this work, we focus on service incompletions in the telehealth setting, which can occur through patient no-shows, i.e., the patient does not go to the clinic at all, or abandonment, i.e., the patient leaves the clinic before being seen by the physicians.

Service incompletion frequently arises in outpatient clinics. From the clinic's perspective, it can result in wasted time and resources, such as staff time and medical supplies. In addition, it can lead to reduced revenue for the healthcare provider (Satiani et al. 2009, Berg et al. 2013). From the patient's perspective, missed appointments can result in delays in care or interruption in adherence to the care plan, which may negatively impact their health (Moore et al. 2001, Schectman et al. 2008). A significant amount of medical and operations management literature has been devoted to understanding patient no-show behavior in clinic appointment scheduling (e.g., Dantas et al. 2018). There has also been a substantial amount of work studying abandonment behavior (also known as

leave without being seen) in the Emergency Department (e.g., Batt and Terwiesch 2015), but this phenomenon has not been studied in the outpatient setting. Understanding service incompletion has two important implications. First, we can design targeted strategies to reduce missed appointments. Second, we can design appropriate scheduling policies to increase throughput and improve service quality (Gupta and Denton 2008). Importantly, if visit incompletion behavior varies depending on visit modality – i.e., in-person versus telemedicine – understanding these differences can be critical to designing well-performing approaches to integrate telemedicine with in-person visits.

In this work, we take an empirical approach to develop a better understanding of service incompletion for in-person and telemedicine appointments. We utilize data from two large internal medicine outpatient clinics at an academic medical center to identify and quantify the two main sources of service incompletion: no-show and abandonment, and how the behavior differs under the in-person and telemedicine visit modalities. We are especially interested in the abandonment behavior since it is not well-studied in the outpatient setting. It is important to differentiate between the two sources of service incompletion since addressing each requires distinct interventions. To mitigate no-shows, clinics can use strategies such as sending reminders in advance, reducing lead time, and providing transportation support for in-person patients. To mitigate abandonment, clinics can implement measures to reduce in-clinic waiting time and provide appropriate information to increase patients' patience. Our empirical findings indicate that telemedicine patients have roughly the same incompletion rate as in-person patients (20.38% and 20.26% for telemedicine and in-person respectively, p-value for two-sided Welch's t-test: 0.5065). However, in-person patients do not appear to be sensitive to intra-day delays, while a physician not being available at the scheduled appointment start time can lead to a 7.40% (95% confidence interval: (3.48%, 11.32%)) increased likelihood of visit incompletion for telemedicine patients.

Managing customers' waiting and abandonment has been a long-standing focus for service providers and is an important topic in operations management (see, e.g., Atar et al. 2010, Allon et al. 2011, Yu et al. 2018). In recent years, with the advancement of technology, there are two new trends in service operations. First, more services are provided online (e.g., shopping, consultation, etc). Some studies have shown that people tend to have shorter time expectations and are more impatient online (Lee et al. 2012, Gallino et al. 2022). Second, the adoption of virtual queues – allowing customers to join a queue

remotely through a mobile application or online portal – has been embraced by many service providers, including restaurants and theme parks. Virtual queues can help reduce perceived waiting time and its associated negative feelings, by offering customers flexibility in the types of activities they can engage in and the locations while waiting. (Hu et al. 2021). Telemedicine visits share similar features with both online services and virtual queues. Thus, it is a priori unclear whether telemedicine patients are more impatient than in-person patients. Our study shows that telemedicine patients are more likely to have incomplete visits when the physician is not available at the appointment start time. We refer to this phenomenon as "abandonment." Several factors could contribute to this phenomenon. One factor could be the sunk cost effect (Ülkü et al. 2020). Note that compared to in-person patients, telemedicine patients may experience a much lower sunk cost due to zero commute requirements. Another factor could be the lack of information when waiting online (Yu et al. 2022), as patients are not able to observe the clinic's operations or progress.

Patients' no-show and abandonment behaviors have important implications on appointment scheduling (see, e.g., Liu and Ziya 2014, Zacharias and Pinedo 2014, Zacharias and Armony 2017). Based on our empirical findings, we conduct counterfactual experiments to analyze how different no-show and abandonment behaviors of telemedicine versus in-person patients affect how these patients should be sequenced within a day. We consider the setting where the provider offers both in-person and telemedicine visits and uses an individual-block-fixed-interval appointment book. Our analysis indicates that it is important to correctly differentiate incompletions due to abandonments from those due to no-shows to determine a good sequencing rule. Not being able to accurately quantify the magnitude of no-show versus abandonment can lead to highly suboptimal sequencing rules in some cases.

While there is extensive research on patient no-shows and their implications on appointment scheduling, to the best of our knowledge, we are the first to study abandonment behavior (i.e., left without being seen) in the outpatient setting. This can be because most previous studies focus on in-person visits only and – as we find – there is little impact of intra-day delay on visit incompletion for in-person patients. Our empirical results and counterfactual experiments indicate that with the wide adoption of telemedicine, it is

important to take both no-show and abandonment behaviors into account when designing scheduling policies.

Beyond its contextual contribution, our work also tackles several econometric challenges. Similar to many studies utilizing retrospectively collected data, we need to properly account for potential endogeneity. In particular, there may be unobserved patient characteristics that are correlated with both physician availability and visit incompletion. In addition to endogeneity, there are also a non-negligible amount of missing values for physician availability, due to missing check-out times. These missing values may not occur randomly, introducing a sample selection bias. Lastly, physician availability can be misreported. The clinic administrative staff may record the check-out time later than the actual service completion time. Thus, a physician may be inaccurately recorded as unavailable (i.e., the previous patient has not checked out yet) while actually being available. To tackle these estimation challenges, we construct appropriate instrumental variables (Wooldridge 2010), model the sample selection via a probit model, i.e., Heckman selection model (Heckman 1979), and model the one-sided misreporting as a partially observable bivariate probit model (Nguimkeu et al. 2019). The combined model is a multivariate probit model which can be estimated via the full information maximum likelihood estimation (Wooldridge 2010). Even though the above three econometric challenges (endogeneity, sample selection, and measurement error) have been tackled individually or in pairs in the literature (see, e.g., Van de Ven and Van Praag 1981, Chan et al. 2017, Nguimkeu et al. 2019), to the best of our knowledge, we are the first to tackle all three simultaneously. We believe these challenges can arise in many other empirical settings, especially in services where human behavior of service providers can result in imperfect data. Thus, our estimation strategies can be applied more broadly.

1.1. Literature Review

We next provide a brief review of the related literature.

Telemedicine. With the growing popularity of telemedicine, there is an increasing number of works studying operational strategies to integrate telemedicine with traditional in-person visits. Rajan et al. (2019) study the impact of patient heterogeneity and telemedicine on the speed-quality trade-off for medical specialists treating chronic conditions. They consider the utility-maximizing behavior of both the specialists and the patients, and found that by lowering the travel burden, telemedicine increases patient

access to specialists. Sun et al. (2020) take an empirical approach to investigate the effect of telemedicine on emergency department (ED) care delivery. They show telemedicine availability in the ED reduces the average patient length of stay without sacrificing care quality or increasing costs for patients. Delana et al. (2022) conduct an empirical study to evaluate the impact of telemedicine centers on physical healthcare-delivery channels in South India. They find that telemedicine increases access and treatment rates. Our work complements this line of research by investigating the no-show and abandonment behaviors of telemedicine patients. Our findings have important implications for the design of appointment scheduling policies to achieve better resource utilization and service quality.

For patient no-shows and customer abandonment, there is a vast amount of literature on different aspects of the problems, such as structural models explaining customers' utility-maximizing behavior, mitigation strategies to reduce no-shows or abandonment, and the implication of these behaviors on scheduling policies. In this section, we focus on empirical works only, which are most related to our work.

Patient no-show. There is extensive literature studying patients' no-show behavior in outpatient clinics. Some of the factors that have been shown to be correlated with no-shows include appointment lead time, patient's prior no-show history, age, socioeconomic status, insurance type, and travel distance (see Dantas et al. 2018 for a review of the medical literature). Kong et al. (2019) identify a significant time-of-day effect on patient show-up probabilities and study appointment scheduling when no-shows are exogenous versus endogenous. Liu et al. (2019) study the effect of rescheduling on patient no-show behavior. Osadchiy and Kc (2017) and Liu et al. (2018) quantify how appointment lead time affects the patient's utility, and thus their decision to show up for the appointment. Most of these works focus on in-person visits only. There are also some recent medical papers that study no-show behavior for telemedicine patients and show that telemedicine is associated with a reduced no-show rate compared to in-person patients (see, e.g., Snoswell and Comans 2021, Bramati et al. 2022, Alkilany et al. 2022). Our work complements these works by studying both no-shows and abandonment – two different sources of service incompletion. We also compare the behavior between in-person and telemedicine patients. To the best of our knowledge, none of the previous works study abandonment behavior due to intra-day delays in the setting of outpatient clinics.

Left without being seen or abandonment. Left without being seen or abandonment, in the healthcare setting, has been almost exclusively studied in the emergency department (Arendt et al. 2003, Batt and Terwiesch 2015, Bolandifar et al. 2019). Customer abandonment or balking is much better understood in other service settings. For example, Mandelbaum and Zeltyn (2013), Akşin et al. (2013), Yu et al. (2017), Hathaway et al. (2022) study abandonment in call centers; Yom-Tov et al. (2021), Gallino et al. (2022) study abandonment in online retailing; Yu et al. (2022) study abandonment in ride-sharing platforms. It is worth noting that in most of the studies in non-healthcare service settings, customers wait online/virtually, and their behaviors can be affected by delay announcements. Our work fills a gap by focusing on outpatient clinics and compares patients' sensitivity to delay when attending medical appointments in-person versus online.

1.2. Paper Organization

The remainder of the paper is organized as follows. Section 2 provides more details about the empirical setting and dataset for our analyses. We introduce our model and discuss several econometric challenges, as well as our approach to addressing them in Section 3. Section 4 presents our empirical findings and sensitivity analysis to demonstrate the robustness of the results. In Section 5, we conduct counterfactual analyses about the implications of our empirical findings on appointment sequencing policies when scheduling a combination of in-person and telemedicine visits. We conclude in Section 6.

2. Setting and Data

We use data from the two largest internal medicine outpatient clinics at Columbia University Irving Medical Center (CUIMC) from Feb 1, 2020 to Dec 31, 2020. The data are at the patient-visit level and contain patient-level information such as age, marital status, and home address, and visit-level information such as visit type (e.g., new patient or follow-up), primary insurance type, and various time-stamps associated with the visit. To complement the main data set, we also extracted hourly weather data in New York City from the National Weather Service by National Oceanic and Atmospheric Administration and median household income for different ZIP Codes from the census data (Rozzi 2021).

Our collaborating clinics are located in Northern Manhattan and offer comprehensive medical care to a broad and diverse population including lower-income, and racial/ethnic minority groups. The service line of internal medicine provides prevention, evaluation,

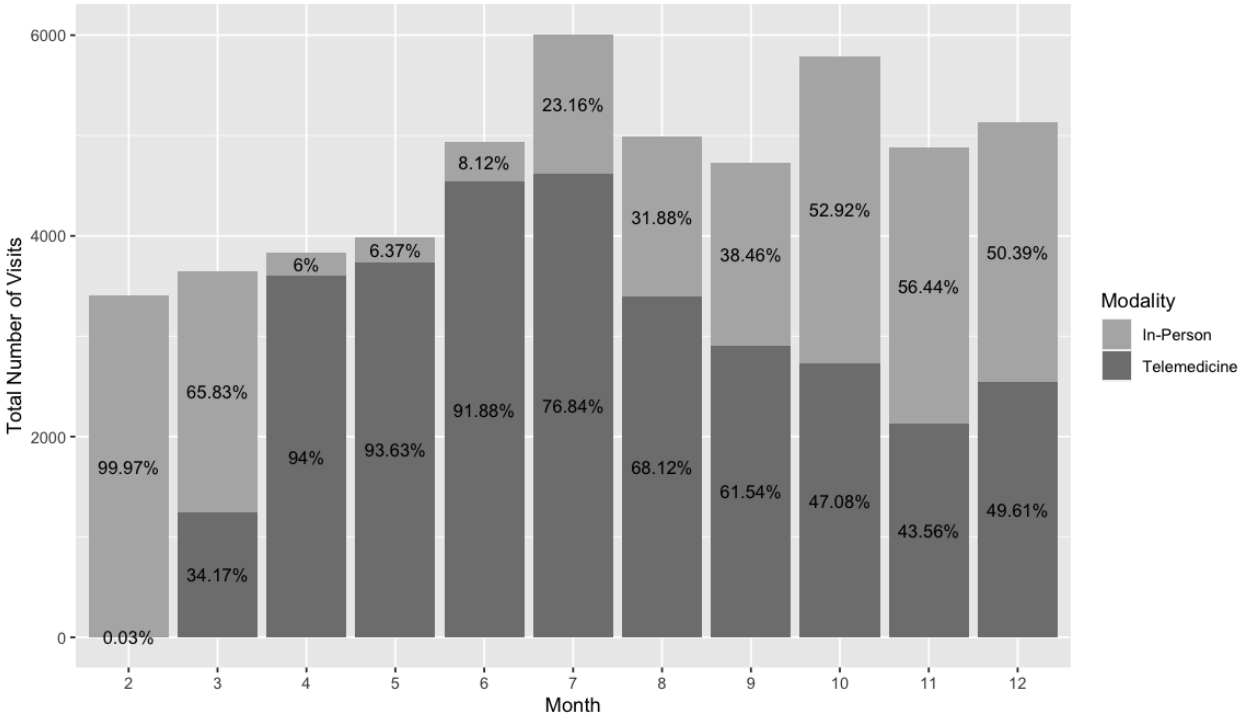


Figure 1 **Stacked bar chart of the total number of visits and the composition of in-person and telemedicine modalities each month during the study period in our collaborating clinics.**

treatment, and management of acute illness and chronic diseases. Before the COVID-19 pandemic (New York State issued the shelter-in-place order on March 20th, 2020), almost all appointments at the clinics were conducted in person. During the pandemic, in response to the surges in COVID-19 cases, the clinics switched to a hybrid mode of in-person and telemedicine visits, where the latter is defined as visits occurring through audio or video equipment. Figure 1 shows the total number of visits and the composition of the two modalities of visits for each month during the study period. We observe that the clinic and patients were able to quickly adapt to telemedicine. In April 2020, telemedicine accounted for 94.00% of total visits. Even after New York City reopened in June 2020, telemedicine still constituted 58.25% of the visits (June to December 2020), suggesting that telemedicine is likely to persist as an integral part of the routine healthcare delivery model.

2.1. Sample Selection

We focus on complete versus incomplete visits. Visits that are canceled in advance are excluded from our study. From the scheduling perspective, clinics have a better chance of re-utilizing these canceled appointment slots. We also exclude appointments that happen during the weekend or outside the 8:00 am to 6:00 pm regular working-hours window.

Self-paid patients or patients with unknown insurance are dropped from our sample as they represent only a small portion of the population and their behavior might be different from those under coverage. Lastly, there are some providers who practice less than 11 days over the 11-month study period; these providers may have been filling in temporarily. We excluded patients of these providers as well. Figure 2 summarizes our data selection processes.

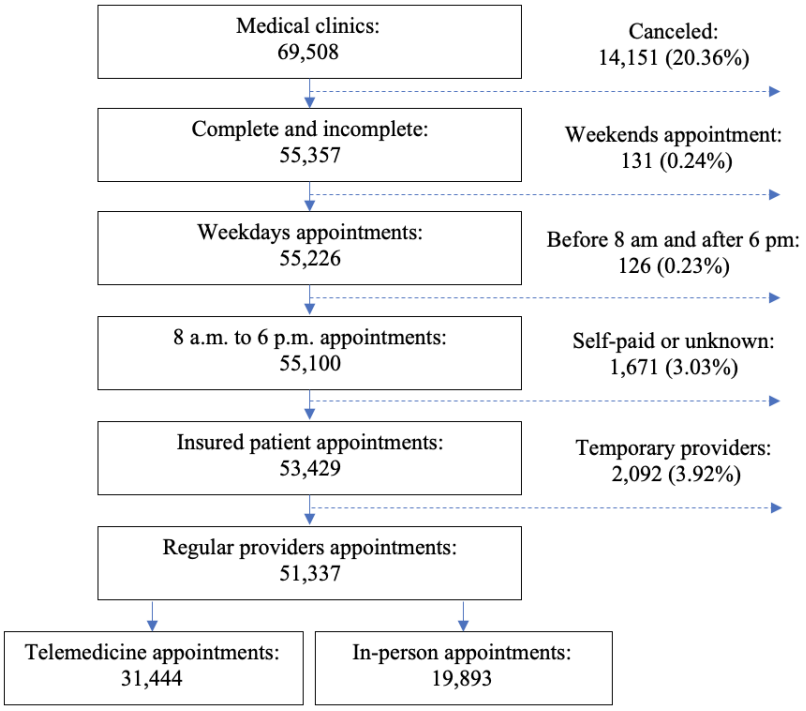


Figure 2 **Selection of the appointment samples.**

The final cohort consists of 51,337 visits from 17,261 unique patients with 249 unique providers. Among them, there are 19,893 in-person visits from 11,593 unique patients with 180 providers, and 31,444 telemedicine visits from 14,052 unique patients with 214 providers.

2.2. Main Outcome Measure: Incomplete Visits

In our data, there are four appointment status categories: "complete", "no show", "left without being seen", and "arrive." If a patient shows up and completes the visit, the appointment status is recorded as "complete." We aggregate "no show", "left without

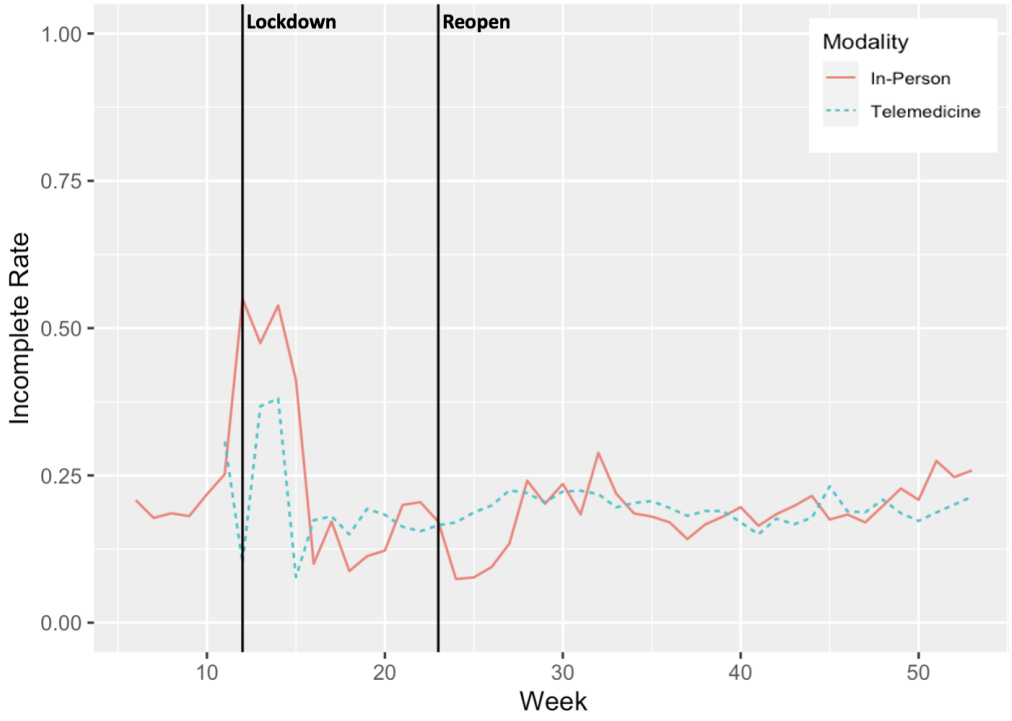


Figure 3 **Time series plots of visit incompletion rate for in-person and telemedicine modalities in different weeks of 2020 in our collaborating clinics.**

being seen", and "arrive" as "incomplete." We define an indicator "Incomplete" to denote the visit status. If a patient shows up but does not complete the visit, the appointment status should be recorded as "left without being seen" or "arrive", but is often recorded as "no show" (0.58% of the no-shows have check-in times but none of them have check-out times). Overall, in our data, there are very few observations in the "left without being seen" or "arrive" categories ("left without being seen" constitutes 0.05% and 0.87% of the in-person and telemedicine appointments respectively, "arrive" constitutes 0.04% and 0.17% of the in-person and telemedicine appointments respectively), and all these observations are missing check-out times.

Figure 3 shows the visit incompletion rate for each week of our study period (week 6 to week 52 of 2020). We note that the incompletion rates for the two modalities are roughly the same over this period, i.e., 20.26% for in-person and 20.38% for telemedicine (p-value for two-sided Welch's t-test: 0.5065). At the beginning of the lockdown period (weeks 12 - 15), we see high incompletion rates for both in-person and telemedicine visits. However, after week 16, the incompletion rates stabilized at around 20%.

2.3. Main Treatment Variable: Physician Availability

We aim to differentiate incomplete visits between *abandonment*, which we assume is driven by the intra-day wait, and *no-show*. To separate the two mechanisms for visit incompletion, we measure the effect of physician availability on "Incomplete." The logic behind our main treatment variable is that if the physician is not available at the appointment start time, patients may experience in-clinic delays and will be more likely to abandon. On the other hand, physician availability is unlikely to affect no-shows.

A physician is defined as being unavailable if there is still a patient under their care. Specifically, we consider all patients whose scheduled start time is within a three-hour window prior to the focal patient's scheduled start time. If any of these patients have a checkout time *after* the focal patient's scheduled start time, the physician is considered unavailable. In the special case where there were no complete visits that began within the three-hour window prior to the focal visit, the physician is considered to be available. For example, for the first appointment with a physician on a day, the physician is considered to be available. Alternatively, if all visits scheduled to start within the three-hour window were incomplete, we would consider the physician to be available. We use an indicator, "Available", to denote the physician availability.

In our data, some check-out times are missing. If we cannot determine the physician availability due to missing check-out times, we assign N/A to "Available." Table 1 summarizes the value of the availability indicator for the in-person and telemedicine cohorts. In many cases, we cannot observe physician availability.

Table 1 Summary statistics of physician availability stratified by modalities of visit

Availability	In-Person	Telemedicine
1	8,204 (41.24%)	10,886 (34.62%)
0	7,371 (37.05%)	6,261 (19.91%)
N/A	4,318 (21.71%)	14,297 (45.47%)

2.4. Control Variables

We also include a number of other variables in our empirical analysis to control for their potential effects on the dependent variable (i.e., Incomplete).

It has been well-documented in the literature that many patient-level and visit-level characteristics can affect patient no-show behavior (Lacy et al. 2004, Dantas et al. 2018).

We control for the following patient-level characteristics: age, gender, race, marital status, household income, and distance from home to the clinic. We define household income as the median annual census income of the patient's home Zip Code (mean 38.26k and standard deviation 13.31k USD per year). We also define the distance from home to the clinic as the distance between the patient's home address to the clinic address on Google Maps; in cases where the accurate distance is not available, we approximate it by the centroid of the Zip Codes (mean 2.90 miles and standard deviation 3.53 miles). At the patient-visit level, we control for the patient's insurance status (commercial, Medicare, or Medicaid), appointment type (new patient, follow-up, etc), lead days, day of the week (Monday to Friday), time of the day (morning or afternoon), rush hours (morning rush: 8:00 - 10:00 am, afternoon rush: 4:00 pm - 6:00 pm, or no rush), lockdown status of New York City (pre-lockdown, lockdown, partially open, open), and precipitation and snow depth for the hour of the scheduled appointment start time. We also include physician-fixed effects and clinic-fixed effects to account for other unobserved characteristics of the physicians and the clinics. See Appendix A for detailed definitions and summary statistics of these control variables.

3. Model and Estimation

Our goal is to estimate the effect of physician availability on visit incompletion. We conduct the estimation for in-person and telemedicine cohorts separately. In this section, we discuss the identification challenges and our identification strategies.

We model the patient-visit level incompletion status as

$$Incomplete_i = \mathbf{1}\{\beta^\top X_i + \delta Available_i + u_i > 0\} \tag{1}$$

Here $Available_i$ indicates whether the focal physician is available at the scheduled appointment start time for visit i. X_i denotes the control variables introduced in Section 2.4, including various patient-level and patient-visit level characteristics, physician-fixed effects, and clinic-fixed effects. u_i is the unobserved determinant of visit i's status.

Our aim is to estimate δ, the treatment effect of physician availability on visit incompletion. Unfortunately, a naive estimate of (1), e.g., via a linear probability model, or a probit or logit model, is likely to suffer from biases due to a number of estimation challenges that arise in our data and setting.

3.1. Identification Challenges

There are three main identification challenges due to the data limitation: endogeneity, missing values, and measurement errors.

3.1.1. Endogeneity Direct estimation of (1) can be biased due to omitted variables that are correlated with both physician availability and visit incompletion. For example, for telemedicine visits, major internet connection problems or other technology failures that last for a period of time can affect many patients' ability to attend their appointments. This will increase both physician availability and visit incompletion rate. For in-person visits, inclement weather and severe traffic congestion (beyond what we have controlled for) can increase both physician availability and visit incompletion rate. Since these unobservables, captured in u_i, are positively correlated with $Available_i$ and $Incomplete_i$, direct estimation of (1) is likely to result in a positive estimation bias of δ.

We apply an instrumental variable (IV) approach to account for the endogeneity (Wooldridge 2010). A valid IV needs to satisfy two conditions: C1) It must be correlated with $Available_i$; C2) It has no direct effect on $Incomplete_i$ other than through $Available_i$, conditional on the other covariates. The IV we use is the number of scheduled visits of the same provider in a three-hour window prior to the focal visit. Note that the physician is more likely to run late for the focal visit if there are many appointments scheduled prior to the focal visits. Thus, the IV is negatively correlated with $Available_i$. Meanwhile, since it is the scheduled appointments (which are scheduled in advance) rather than complete visits, the IV is unlikely to be correlated with unobservable determinants of visit status during the day of the appointment. Note that in our clinic, same-day appointments only account for 1.07% of the total visits.

Since we control for physician-fixed effects, the main potential mechanism that may affect the exclusion restriction is that there can be unobserved patient characteristics that are correlated with both the scheduling decision and visit incompletion. For example, patients of different severity levels tend to have different incompletion rates; more severe patients may be more likely to follow the care plan and attend visits. Meanwhile, the staff may take the severity information into account when scheduling patients. Even though internal medicine focuses on primary care rather than urgent care, and we control for the appointment type, there can still be unobserved patient severity information that affects the validity of the IV. In Appendix B, we run additional regression analyses using variables

that are likely to be correlated with patient severity (e.g., number of historical visits) to show that in our clinics, patient severity is unlikely to be correlated with the IV (see Table 7). In addition, it has been shown in the literature that patients' no-show history is correlated with their no-show behavior for the current appointment (Dantas et al. 2018), and some clinics may utilize patients' no-show history in scheduling (e.g., overbooking) (Samorani et al. 2022). First, we do not see much overbooking in our data. Second, in Appendix B, we run additional regression analyses to show that no-show history is unlikely to have a significant effect on how appointments are scheduled in our clinics (see Table 8). Lastly, we run a sensitivity analysis of the exclusion restriction in Section 4.3, and show that a mild violation of the exclusion restriction is unlikely to affect our estimation results.

3.1.2. Missing Values Due to missing check-out times, our main treatment variable, physician availability, has a significant amount of missing values (22% for in-person, 45% for telemedicine). These missing check-out times may not be missing at random. In particular, since most of these time stamps are recorded by the clinic's administrative staff when they check out the patients, we hypothesize that when the clinic is busy, it is more likely to have missing check-out times.

To account for the potential sample selection bias, following the Heckman selection model (Heckman 1979), we model the sample selection process via a probit model. To ensure estimation stability, in the probit model, it is desirable to have a variable that is correlated with the selection probability but uncorrelated with visit incompletion. For this, we use the number of scheduled visits of the same clinic excluding the focal provider in a three-hour window around (1.5 hours before and 1.5 hours after) the focal visit, which is a measure of clinic busyness level. Since administrative staff is typically shared among different providers in the same clinic, we expect that when the clinic is busy, it is more likely to have missing check-out times. Meanwhile, since we exclude the number of scheduled visits of the focal provider, this variable is unlikely to affect the visit incompletion of the focal visit (see Appendix B Table 9).

3.1.3. Measurement Errors Even when the check-out times are not missing, they may be recorded with errors. In particular, the check-out time can be recorded much later than when the patient finishes service. In a few extreme cases, we observe an unusual number of patients being checked out during a small time window towards the end of the day. These

recording errors appear to occur more often when the clinic is busy, i.e., when the clinic's front-office staff are busy with many other demanding duties.

The wrongly recorded check-out times lead to measurement errors when we calculate physician availability. In particular, the measurement error happens when a physician is available for the focal visit but is recorded as unavailable because the check-out times of the previous patients are recorded much later than when they actually finish service.

Following Nguimkeu et al. (2019), we model the one-sided misreporting as a partially observable bivariate probit model. In particular, we have two probit models: the first one models whether the physician is available; the second one models whether the availability is correctly reported. The observed physician availability is the product of the two probit models. To ensure estimation stability, in the second probit model, it is desirable to have a variable that is correlated with the recording error but uncorrelated with visit incompletion. For this, we use the ratio between the number of scheduled visits to the same clinic excluding the focal provider in a three-hour window prior to the focal visit and the total number of scheduled visits to the same clinic excluding the focal provider in a three-hour window prior to and a three-hour window after the focal visit. The idea is that if the clinic is busier before the focal visit than after, the staff are more likely to record the check-out time of the previous patients at a later time.

3.2. Main Model

In this section, we introduce our main model that addresses the three identification challenges discussed in Section 3.1. The model contains four parts representing i) whether the visit is incomplete, ii) whether the physician is available, iii) whether the physician's availability is reported accurately, and iv) whether the physician's availability is observed.

$$Incomplete_i = \mathbf{1}\{\beta_1^\top X_{1i} + \delta Available_i + u_{1i} > 0\} \tag{2}$$

$$Available_i = \mathbf{1}\{\beta_2^\top X_{2i} + \alpha_2 PreWorkProvider_i + u_{2i} > 0\} \tag{3}$$

$$ReportTrue_i = \mathbf{1}\{\beta_3^\top X_{3i} + \alpha_3 RelWorkClinic_i + u_{3i} > 0\} \tag{4}$$

$$Observe_i = \mathbf{1}\{\beta_4^\top X_{4i} + \alpha_4 WorkClinic_i + u_{4i} > 0\} \tag{5}$$

Note that in our data, if $Observe_i = 1$, we are only able to observe $Available_i \times ReportTrue_i$, which we denote as AR_i. We never directly observe $Available_i$ and $ReportTrue_i$ individually.

In equation (2), X_{1i} is a vector of control variables containing patient-level and patient-visit level characteristics, physician-fixed effects, and clinic-fixed effects (see Section 2.4 for more details). $Available_i$ denotes the **true** availability of the physician (which we do not directly observe). In equation (3), X_{2i} is the same as X_{1i}. $PreWorkProvider_i$ is the IV, which is the number of scheduled visits of the same provider in a three-hour window prior to the focal visit. In equation (4), X_{3i} is the same as X_{1i}. $RelWorkClinic_i$ is the clinic's relative workload, which is a ratio defined as follows. Let N_{pre} denote the number of scheduled visits of the same clinic excluding the focal provider in a three-hour window prior to the focal visit's scheduled start time. Let N_{post} denote the number of scheduled visits of the same clinic excluding the focal provider in a three-hour window after the focal visit's scheduled start time. Then, $RelWorkClinic_i = N_{pre}/(N_{pre}+N_{post})$. In equation (5), X_{4i} contains all the covariates in X_{3i}, and also $PreWorkProvider_i$ and $RelWorkClinic_i$. $WorkClinic_i$ is the number of scheduled visits of the same clinic excluding the focal provider in a three-hour window around the focal visit's scheduled start time (1.5 hours before and 1.5 hours after). $u_{1i}, u_{2i}, u_{3i}, u_{4i}$ are the unobserved determinates which are assumed to follow a multivariate Normal distribution:

$$(u_{1i}, u_{2i}, u_{3i}, u_{4i})|X_i \sim N(0,\Sigma) \text{ where } X_i = (X_{4i}^\top, WorkClinic_i)^\top \text{ and } \Sigma = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 \end{pmatrix}.$$

In particular, the model is a multivariate probit model which can be jointly estimated via the full information maximum likelihood estimation (Wooldridge 2010).

Under suitable regularity conditions, we can establish the consistency and asymptotic normality of the estimator. Denote the data by z_i, i.e.,

$$z_i = (X_i^\top, Incomplete_i, AR_i, Observe_i)^\top$$

when $Observe_i = 1$, and

$$z_i = (X_i^\top, Incomplete_i, Observe_i)^\top$$

when $Observe_i = 0$. We also denote the parameters by

$$\theta = (\beta_1^\top, \delta, \beta_2^\top, \alpha_2, \beta_3^\top, \alpha_3, \beta_4^\top, \alpha_4, \rho_{12}, \cdots, \rho_{34})^\top.$$

Let $f(z;\theta)$ denote the density of z under parameter θ, and θ_0 denote the true value of the parameter. We also denote Θ as a compact set for the parameters with $\theta_0 \in \Theta$, i.e., the likelihood function is optimized in Θ. Lastly, we write $\hat{\theta}_n$ as the full information maximum likelihood estimator with n observations.

PROPOSITION 1. *Suppose* $(z_1, z_2, \cdots, z_n)$ *are independent, identically distributed, and bounded. Assume (i)* $\mathbb{E}[X_i X_i^\top]$ *is nonsingular; (ii)* $\mathbb{E}[(\nabla_\theta \ln f(z_i;\theta_0))(\nabla_\theta \ln f(z_i;\theta_0))^\top]$, *denoted by* I_{θ_0}, *is nonsingular; (iii)* $\theta_0 \in interior(\Theta)$, *then*

$$\hat{\theta}_n \xrightarrow{p} \theta_0 \ as \ n \to \infty$$

and

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I_{\theta_0}^{-1}) \ as \ n \to \infty.$$

The proof of Proposition 1 can be found in Appendix D.

4. Estimation Results

In this section, we present our estimated effects of physician availability on the visit incompletion rate. Table 2 summarizes the estimation results based on the main model (2) – (5) (Full Model). To facilitate interpretation, we report the average marginal effect (Greene 2003), which is defined as

$$\frac{1}{N}\sum_{i=1}^{N}(\Phi(\hat{\beta}_1^\top X_{1i} + \hat{\delta}) - \Phi(\hat{\beta}_1^\top X_{1i}))$$

for binary variables such as *Available* and

$$\hat{\alpha}_2 \frac{1}{N}\sum_{i=1}^{N}\phi(\hat{\beta}_2^\top X_{2i} + \hat{\alpha}_2 PreWorkProvider_i)$$

for continuous variables such as $PreWorkProvider$, where $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\delta}$, and $\hat{\alpha}_2$ are the estimated coefficients. $\Phi(\cdot)$ and $\phi(\cdot)$ denote the cumulative distribution function and probability density function of the standard Normal distribution respectively. We also present the result from a direct estimation of (1) (Naive Model), ignoring endogeneity, missing values, and measurement errors, for comparison.

The full model estimation shows that physician availability does not have a significant impact on the in-person incompletion rate but has a significant negative effect on the telemedicine incompletion rate. The average marginal effect of physician availability on

Table 2 Naive model versus full model estimates of the impact of physician availability. We report the estimated average marginal effect and the corresponding standard error (in parentheses)

	Naive Model		Full Model	
	In-Person	Telemedicine	In-Person	Telemedicine
Incomplete				
Available	0.033^{***}	0.004	0.016	-0.074^{***}
	(0.007)	(0.007)	(0.019)	(0.020)
Available				
PreWorkProvider			-0.086^{***}	-0.089^{***}
			(0.002)	(0.002)
ReportTrue				
RelWorkClinic			-0.446^{***}	-0.485^{***}
			(0.016)	(0.019)
Observe				
WorkClinic			-0.001^{***}	-0.002^{***}
			(0.000)	(0.000)
ρ_{12}			0.180^{***}	0.191^{***}
ρ_{13}			0.010	0.006
ρ_{14}			-0.005	0.065
N	15575	17147	19893	31444

Note: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

telemedicine incompletion rate is -7.4%, i.e., if the physician is available at the scheduled start time of the focal telemedicine visit, the incompletion rate decreases by 7.4%. We conjecture that the effect of physician availability on the visit incompletion rate is due to abandonment, i.e., patients get impatient when waiting for the doctor and leave without being seen. This is because no-shows are unlikely to be affected by real-time physician availability as the patient did not even make it to the clinic.

Define the abandonment rate as the incompletion rate conditional on the patient showing up and the provider not being available at the scheduled start time. Let γ denote the abandonment rate. We further denote η as the incompletion rate, p as the no-show rate, and k as the physician availability rate. Then, we have

$$\eta = p + (1-p)(1-k)\gamma,$$

Along with the fact that the treatment effect δ can be written as the difference between the incompletion rate conditional on physician available (i.e., p) and the incompletion rate conditional on physician unavailable (i.e., $p+(1-p)\gamma$), we can solve for the unknowns γ and p. In particular,

$$\gamma = \frac{-\delta}{1-\eta-\delta(1-k)}.$$

For telemedicine patients, plugging in the empirical visit incompletion rate and physician availability rate, and our estimated $\hat{\delta}$, the estimated abandonment rate is 9.0% (95% confidence interval: (4.39%, 13.61%)).

Based on the above equation, we can also estimate the no-show rate, which is 20.26% (95% confidence interval: (18.40%, 22.12%)) for in-person patients and 17.68% (95% confidence interval: (16.25%, 19.11%)) for telemedicine patients. This indicates that telemedicine patients have a slightly lower no-show rate than in-person patients (p-value for two-tailed test: 0.03), which is consistent with previous observations in the literature (Franciosi et al. 2021).

Our estimated treatment effect suggests that telemedicine patients are more sensitive to intra-day delays than in-person patients. This may look surprising at first glance since previous studies suggest that waiting virtually is more pleasant than waiting in person in many service settings (Hu et al. 2021). However, thinking more carefully about the setting of an outpatient clinic, several factors could contribute to what we observe. First, by physically being in the clinic, in-person patients spend time and money traveling. There can be a sunk cost effect that makes them more willing to wait (Sweis et al. 2018). Second, when waiting in the clinic, in-person patients are able to gather more information about how busy the clinic is. The operational transparency and information availability could also make them more willing to wait (Hui and Tse 1996, Buell et al. 2017, Ansari et al. 2022, Park et al. 2022). Lastly, people may have a higher expectation to be on time when meeting online (Lee et al. 2012). We also acknowledge the possibility that the more severe patients may be more likely to choose in-person visits and these patients may also be less likely to abandon.

By comparing the estimation results based on the naive model versus the full model, we note that ignoring the identification challenges leads to biased estimates. When we directly estimate the naive model, physician availability has a positive and significant effect on the in-person incompletion rate but does not have a significant impact on the telemedicine incompletion rate. This is in contrast to the estimates based on the full model. The positive estimation bias due to endogeneity is well expected as discussed in Section 3.1.1. The endogeneity issue can also be seen from that ρ_{12}, the correlation between μ_{1i} and μ_{2i}, is positive and significantly different from zero.

We also note that the coefficients for $PreWorkProvider$, $RelWorkClinic$, and $WorkClinic$ in equations (3), (4), and (5) respectively are negative and significantly different from zero, suggesting that they satisfy the relevance conditions, respectively. In particular, the busier the physician is, the less likely that the physician is available at the scheduled start time for the focal patient. The busier the clinic is, the less likely it is to have recorded or accurately recorded check-out times. Lastly, ρ_{14} is not significantly different from zero, suggesting that sample selection is unlikely to significantly bias our estimation.

4.1. Estimation Using Post Lockdown Data Only

In this section, we estimate the main model (2) – (5) using data from June to December 2020 only. New York City started to reopen on June 8 after a nearly three-month lockdown. Both the daily reported new cases and the number of deaths from COVID-19 in New York City remained quite low during this period of time (e.g., the average number of new cases per day was below 1000). Note that this period is before the spread of the delta and omicron variants.

In addition, telemedicine had been adopted for more than three months by June. We assume the physicians and patients are more familiar with this new modality of visits during this time frame. We observe in Figures 1 and 3 that the visit volume and incompletion rates are relatively stable for this period of time.

The estimation results are summarized in Table 3, which are consistent with our estimates using the full data in Table 2. In particular, physician availability reduces the telemedicine incompletion rate by -7.2% on average but has no significant impact on the in-person incompletion rate.

4.2. Alternative Model and Variable Specifications

We explore several alternative model specifications, which we refer to as reduced models. We consider three reduced models. In the first one, we ignore the measurement error but take endogeneity and sample selection into account (IV + Selection), and consider a linear model for $Incomplete_i$ and $Available_i$ (see a similar model in Chan et al. 2017):

$$\begin{aligned} Incomplete_i &= \beta_1^\top X_{1i} + \delta Available_i + u_{1i} \\ Available_i &= \beta_2^\top X_{2i} + \alpha_2 PreWorkProvider_i + u_{2i} \\ Observe_i &= \mathbf{1}\{\beta_4^\top X_{4i} + \alpha_4 WorkClinic_i + u_{4i} > 0\} \end{aligned}$$

Table 3 Naive model versus full model estimates of the impact of physician availability for patient visits in June - December 2020. We report the estimated average marginal effect and the corresponding standard error (in parentheses)

	Naive Model		Full Model	
	In-Person	Telemedicine	In-Person	Telemedicine
Incomplete				
Available	0.040^{***}	-0.008	0.038	-0.072^{***}
	(0.008)	(0.007)	(0.022)	(0.022)
Available				
PreWorkProvider			-0.128^{***}	-0.106^{***}
			(0.002)	(0.002)
ReportTrue				
RelWorkClinic			-0.690^{***}	-0.627^{***}
			(0.031)	(0.028)
Observe				
WorkClinic			-0.002^{***}	-0.003^{***}
			(0.000)	(0.000)
ρ_{12}			0.101^{**}	0.140^{**}
ρ_{13}			-0.086	0.138^{*}
ρ_{14}			-0.099	0.005
N	9803	13467	13604	22859

Note: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

In the second one, we ignore the sample selection but take endogeneity and measurement error into account (IV+Error), and consider a linear model for $Incomplete_i$ (see a similar model in Nguimkeu et al. 2019):

$$Incomplete_i = \beta_1^\top X_{1i} + \delta Available_i + u_{1i} \tag{6}$$

$$Available_i = \mathbf{1}\{\beta_2^\top X_{2i} + \alpha_2 PreWorkProvider_i + u_{2i} > 0\} \tag{7}$$

$$ReportTrue_i = \mathbf{1}\{\beta_3^\top X_{3i} + \alpha_3 RelWorkClinic_i + u_{3i} > 0\} \tag{8}$$

In the third one, we only take endogeneity into account (IV) and consider the following linear models:

$$Incomplete_i = \beta_1^\top X_{1i} + \delta Available_i + u_{1i}$$

$$Available_i = \beta_2^\top X_{2i} + \alpha_2 PreWorkProvider_i + u_{2i}$$

The results are summarized in Table 4. We observe that similar to our main model (Full Model) estimation, in all these alternative models, physician availability does not have a significant impact on the in-person incompletion rate but has a negative and significant

impact on the telemedicine incompletion rate. The average marginal effect of physician availability on telemedicine incompletion rate ranges from -7.4% to -13.4% (95% confidence intervals: (4.58%, 22.22%) for IV+Selection Model, (2.30%, 12.50%) for IV+Error Model, (4.44%, 18.56%) for IV Model). Among the three models, IV+Error provides the closest estimates to our main model.

We also explore alternative specifications of the treatment variable. In our main estimation, physician availability is defined based on whether there is still a previous patient under the physician's care at the scheduled start time of the focal visit. We also try defining a physician as being unavailable if there is still a patient under their care 5/10/15 minutes after the scheduled start time of the focal visit, and re-estimate the main model (2) – (5) (Full Model). The results are summarized in Appendix C Table 10. The average marginal effect of physician availability on telemedicine incompletion rate ranges from -7.0% to -7.3% (95% confidence intervals: (3.38%, 11.22%) for +5min Model, (3.38%, 10.82%) for +10min Model, (2.69%, 11.31%) for +15min Model), and physician availability does not have a significant impact on the in-person visit incompletion rate. These results are consistent with our main estimation.

Table 4 Reduced models (IV+Selection ignoring measurement errors, IV+Error ignoring missing values, IV ignoring both missing values and measurement errors) estimates of the impact of physician availability. IMR is the Inverse Mills Ratio in the Heckman selection model. For linear models, we report the estimated coefficient, for probit models, we report the estimated average marginal effect. The corresponding standard errors are in parentheses

	IV+Selection		IV+Error		IV	
	In-Person	Telemedicine	In-Person	Telemedicine	In-Person	Telemedicine
Incomplete						
Available	0.034	−0.134**	0.020	−0.074**	0.030	−0.115***
	(0.022)	(0.045)	(0.018)	(0.026)	(0.019)	(0.036)
IMR	0.009	−0.016				
	(0.028)	(0.036)				
Available						
$PreWorkProvider$	−0.109***	−0.095***	−0.112***	−0.081***	−0.107***	−0.086***
	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)
ReportTrue						
$RelWorkClinic$			−0.728***	−0.563***		
			(0.023)	(0.025)		
Observe						
$WorkClinic$	−0.001***	−0.002***				
	(0.000)	(0.000)				
N	19893	31444	15575	17147	15575	17147

Note: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

4.3. Sensitivity Analysis of the IV

For the exclusion restriction of the IV, the main concern is that the clinic may use unobserved patient characteristics in scheduling (Note that the unobserved physician or clinic characteristics are controlled for in our model through fixed effects). In this section, we conduct a sensitivity analysis of the IV regarding potential violations of the exclusion restriction using the IV+Error model (6) – (8). We use the reduced model rather than the full model since, based on the analysis in Section 4.2, sample selection does not appear to have a significant impact on our estimation results, and the reduced model greatly decreases the computational complexity of the sensitivity analysis.

Following Baiocchi et al. (2014), suppose there is an unobserved confounder v_i, with mean 0 and variance 1, that is correlated with the outcome $Incomplete_i$ and the IV $PreWorkProvider_i$, but is uncorrelated with the other measured covariates X_{1i}. We assume the linear model

$$Incomplete_i = \beta_1^\top X_{1i} + \delta Available_i + \psi v_i + u_{1i}$$
$$v_i = \xi PreWorkProvider_i + \epsilon_i,$$

with

$$\mathbb{E}[\epsilon_i | X_{1i}, Available_i, PreWorkProvider_i] = 0$$
$$\mathbb{E}[u_{1i} | X_{1i}, Available_i, PreWorkProvider_i] = 0.$$

Note that ψ and ξ are sensitivity parameters where the former measures the effect of a one standard deviation change in the unobserved confounder on the visit incompletion rate and the latter measures the effect of one unit of change in the IV on the unobserved confounder (in the unit of standard deviation).

Recall that the availability we observe in the dataset is $Available_i \times ReportTure_i$. The model (6) – (8) is studied in Nguimkeu et al. (2019) and the authors propose a two-step estimation that provides a consistent and asymptotically normal estimate of δ. In particular, in the first step, α_2, α_3, β_2, and β_3 can be estimated from the partially observable model:

$$\begin{aligned} &Available_i \times ReportTrue_i \\ =&\mathbf{1}(\beta_2^\top X_{2i} + \alpha_2 PreWorkProvider_i + u_{2i} > 0, \beta_3^\top X_{3i} + \alpha_3 RelWorkClinic_i + u_{3i} > 0). \end{aligned}$$

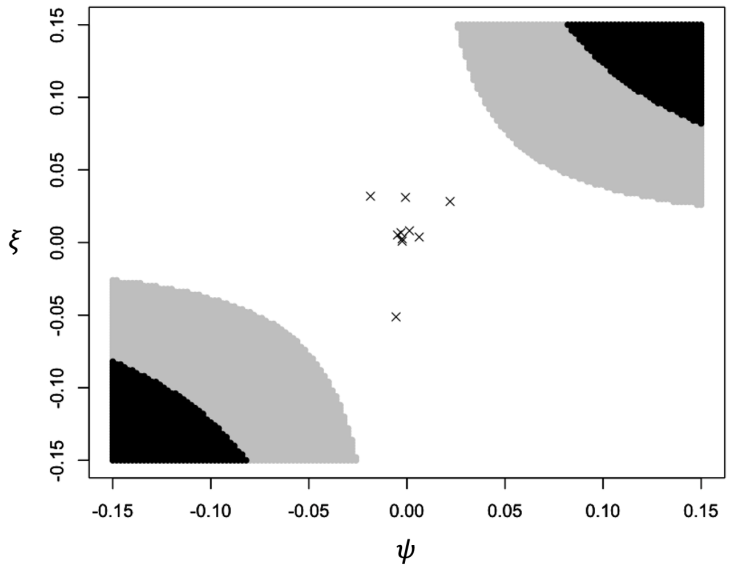


Figure 4 **Sensitivity analysis based on the reduced model (IV+Error). The (ψ, ξ)-values for all observed covariates depicted by crosses are all within the white region, where the estimates for δ are negative and significant at the 5% level. The grey (black) regime depicts where the estimates for δ are insignificant (positive and significant).**

In the second step, $Available_i$ is substituted by the predicted probability of being available

$$\Phi\left(\hat{\beta}_2^\top X_{2i} + \hat{\alpha}_2 PreWorkProvider_i\right)$$

in (6). For the sensitivity analysis, we can construct a consistent estimate for δ using the same two-step estimation by treating $Incomplete_i - \psi\xi PreWorkProvider_i$ as the outcome variable in the second step, i.e.,

$$Incomplete_i - \psi\xi PreWorkProvider_i = \beta_1^\top X_{1i} + \delta Available_i + \psi\epsilon_i + u_{1i}$$

Note that the error term $\psi\epsilon_i + u_{1i}$ has mean zero conditional on X_{1i} and $PreWorkProvider_i$. We run regressions for different values of ψ and ξ to determine parameter regimes where the estimate of δ is a) negative and statistically significant at the 5% level, b) positive and statistically significant at the 5% level, or c) statistically not different than 0 at the 5% level. If the regime where a) holds is large, we are more confident that our results are robust to potential violations of the exclusion criteria.

Figure 4 summarizes the sensitivity analysis result. The white region depicts the area where the estimate for δ is negative and statistically significant at the 5% level. The grey region is where the estimate for δ is statistically not different than 0, and the black region is

where the estimate for δ is positive and statistically significant. We observe that the white area is quite large and the (ψ, ξ)-values for the observed covariates (depicted by crosses) are all within the white region. This indicates that in order for there to be an unobserved confounder that would explain away our result, the effect size of this unobserved confounder would have to be much larger than that of any observed covariates.

5. Counterfactual Analysis on Appointment Sequencing

Our empirical analysis shows that physicians being unavailable at the scheduled appointment start time can lead to a higher rate of incomplete service for telemedicine patients, but it does not have a significant effect on in-person visits. This implies that telemedicine service incompletion can be due to two sources: abandonment – after arriving to the appointment – and no-show – never arriving to the appointment, while in-person service incompletion is mainly due to no-shows. In this section, we study the implications of this empirical finding on the appointment sequencing decision.

There are many aspects of appointment scheduling, such as panel size design, overbooking strategies, and slot-size assignment that could be impacted by no-show and abandonment behavior (Cayirli and Veral 2003). In order to highlight the importance of differentiating between abandonment and no-show behavior, we focus on one aspect of appointment management – the sequencing decision – while keeping the panel size and slot size fixed. Given the heterogeneous impact of prior appointments on the incompletion likelihood of the focal patient, one could expect the sequencing decision to greatly impact visit incompletion as well as provider overtime. We leave a more comprehensive study that considers all aspects of appointment scheduling and accounts for the heterogenous incompletion behaviors of in-person and telemedicine patients for future research.

Let p_{tel} denote the no-show rate for telemedicine patients and p_{in} denote the no-show rate for in-person patients. In particular, we assume each telemedicine (in-person) patient has a probability p_{tel} (p_{in}) of not showing up, independent of all other patients. We also denote γ as the abandonment rate for telemedicine patients. If the physician is busy serving other patients at the scheduled appointment start time of the focal telemedicine patient, the patient will leave without being seen with probability γ. We assume in-person patients have zero abandonment probability, based on our empirical findings. Note that in order to incorporate our empirical findings into our model, we explicitly differentiate the no-show and abandonment rates for telemedicine and in-person patients.

Focusing on the *sequencing* decision, we consider an individual-block-fixed-interval appointment book. We assume the service times are independent and identically distributed exponential random variables whose average is normalized to one unit of time. Each appointment day for the physician has T units of time. Let N denote the total number of patients scheduled for the day. We allow $N \geq T$ at various levels, which corresponds to different overbooking strategies. Each patient is assigned a time slot of length T/N. In particular, let t_i denote the scheduled start time of the i-th patient. Then, $t_1 = 0$, i.e., the start of the day is set as time zero, and $t_i = t_{i-1} + T/N$ for $i = 1, 2, \ldots, N$. We assume patients who show up are punctual. We assume 50% of the visits are telemedicine, and the remaining are in-person. Lastly, we assume patients to be scheduled are known or can be optimally selected by the scheduler in advance, which is a common assumption in prior literature (e.g., Cayirli and Veral 2003, Kong et al. 2019). The decision we consider is how to sequence in-person versus telemedicine patients within the same day. This decision has been studied in the scheduling and sequencing literature with similar assumptions as we made above (e.g, Zacharias and Pinedo 2014, Zacharias and Armony 2017, Zacharias et al. 2022), but in the absence of patient abandonment.

Consistent with the intraday sequencing literature which aims to optimize the trade-off between patient waiting and capacity utilization, our objective is to minimize a combination of in-clinic waiting time, the number of abandoned patients, and physician overtime. The abandonment cost is normalized to be 1 and we vary the values of the waiting cost and overtime cost in our analysis. In particular, the objective function takes the form

$$C_\pi(p_{tel}, p_{in}, \gamma) = Abandon_\pi(p_{tel}, p_{in}, \gamma) + c_w Waiting_\pi(p_{tel}, p_{in}, \gamma) + c_o Overtime_\pi(p_{tel}, p_{in}, \gamma)$$

where π is the sequencing policy, $Abandon_\pi(p_{tel}, p_{in}, \gamma)$ is the average number of abandoning (telemedicine) patients, $Waiting_\pi(p_{tel}, p_{in}, \gamma)$ is the average total waiting time among all patients, $Overtime_\pi(p_{tel}, p_{in}, \gamma)$ is the average overtime of the physician per day.

An optimal sequencing policy is likely to be dependent on the distribution of patient no-show and abandonment behavior as well as cost parameters. The black-box nature of an optimal policy can make it difficult to translate into clinical practice. In practice, clinics prefer to use simple, easy-to-implement approaches to scheduling. As such, we consider four heuristic sequencing rules, based on feedback from our clinical providers:

1. **Block In-Person First** rule schedules in-person patients in the beginning (as a block) and all telemedicine patients after all of the in-person visits.
2. **Block Telemedicine First** rule schedules telemedicine patients in the beginning (as a block) and all in-person patients after all of the telemedicine visits.
3. **Alternating In-Person First** rule alternates between in-person and telemedicine patients and starts with an in-person patient.
4. **Alternating Telemedicine First** rule alternates between telemedicine and in-person patients and starts with a telemedicine patient.

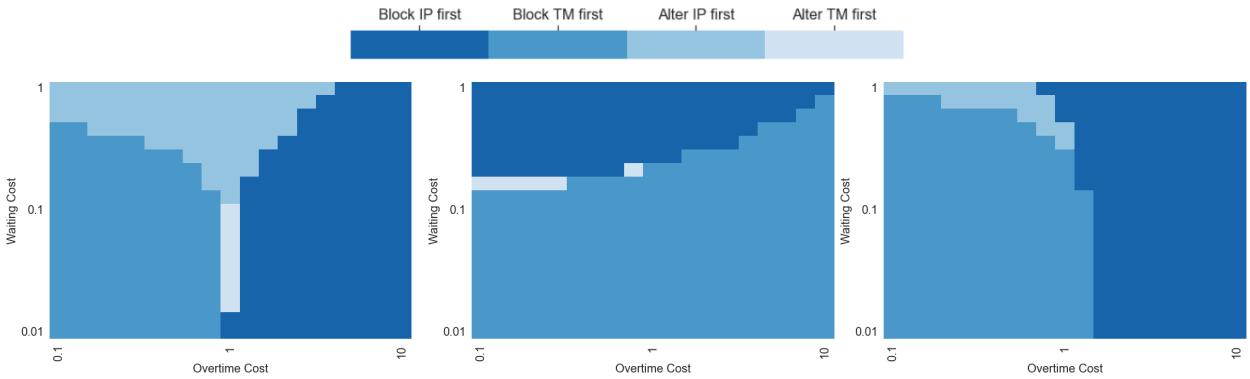
We study which heuristic policy performs the best when in-person and telemedicine patients have different no-show and abandonment behavior. We are especially interested in how patient abandonment affects the sequencing decision.

5.1. Best Heuristic Sequencing Policy

We first study the best heuristic policy under different no-show and abandonment rates, and different waiting and overtime costs. Since the waiting cost is, in general, smaller than the overtime cost (Cayirli et al. 2008), we vary c_w from 0.01 to 1, and c_o from 0.1 to 10.

One key observation from scheduling literature that will arise in our experiments is that patients who are scheduled later in the day are more likely to experience delays than patients who are scheduled earlier in the day (Hassin and Mendel 2008, Zhang et al. 2022).

We start by considering a setting that closely resembles our clinics where $p_{in} = 0.2$, $p_{tel} = 0.18$, and $\gamma = 0.09$. The best-performing heuristic sequencing rule under different waiting and overtime costs is summarized in Figure 5(a). We observe that if the overtime cost dominates other costs (i.e., $c_o > 1$, $c_w < 0.1$), Block In-Person First policy performs the best. The rationale is that telemedicine patients have a higher chance of experiencing delays if they are scheduled later in the day. If the physician is running late towards the end of the day, more telemedicine patients will abandon, leading to an overall higher incompletion rate than in-person patients, resulting in a lower risk of running overtime. If the abandonment cost outweighs the other costs (i.e., $c_o < 1$, $c_w < 0.1$), Block Telemedicine First policy performs the best. This is because when scheduled at the beginning of the day, telemedicine patients have a lower chance of experiencing delays and, subsequently, a lower risk of abandonment. We also note from our numerical experiments that when c_w is large (i.e., $c_w > 0.1$), even though the best-performing policy may not necessarily be Block In-Person First for large overtime cost or Block Telemedicine First for large abandonment



(a) $p_{tel} = 18\%$, $\gamma = 9\%$ (b) $p_{tel} = 10\%$, $\gamma = 9\%$ (c) $p_{tel} = 10\%$, $\gamma = 18\%$

Figure 5 **Best heuristic sequencing policies for different values of p_{tel} and γ, and under different cost parameters.** $p_{in} = 20\%$, $T = 30$, $N = 36$

cost, the optimality gap between the corresponding block scheduling policy and the best-performing policy, Alternating In-Person First, is very small – less than 0.2%.

Prior studies have shown that telemedicine has the potential to substantially reduce the no-show rate by lowering barriers to attending medical appointments (i.e., lowering time commitment and travel costs) (Wootton et al. 2011, Russo et al. 2016, Paquette and Lin 2019). Thus, we also consider a scenario where the telemedicine no-show rate is much lower than the in-person no-show rate, i.e., $p_{in} = 20\%$ while $p_{tel} = 10\%$. We test a low and a high abandonment rate, i.e., $\gamma = 9\%$ versus 18%. Figures 5(b) and (c) summarize the best-performing heuristic policy under different costs. We note that when $\gamma = 9\%$, even when all telemedicine patients are delayed, the overall incompletion rate, $10\% + (1 - 10\%)9\% = 18.1\%$, is still lower than the no-show rate of in-person patients. In this case, Block Telemedicine First performs the best for both low and high overtime costs. We comment that when c_w is large, even though the optimal policy is not Block Telemedicine First, the optimality gap between the corresponding best-performing policy and Block Telemedicine First is small, i.e., at most 1%. When γ is large, i.e., $\gamma = 18\%$, if many telemedicine patients are delayed, the overall service incompletion rate can be higher than the no-show rate of in-person patients. Thus, similar to the case studied in Figure 5(a), when the overtime cost is large enough ($c_o > 2$), Block In-Person First performs the best. When we care more about the abandonment cost, Block Telemedicine First performs the best.

5.2. Abandonment versus No-Show

We next take a closer look into the two sources of service incompletion and analyze the importance of accurately quantifying the abandonment versus no-show rates. Recall that the telemedicine incompletion rate η_{tel} can be decomposed as

$$\eta_{tel} = p_{tel} + (1 - p_{tel})(1 - k)\gamma,$$

where k is the proportion of time the physician is available at the scheduled appointment start time. In what follows, we set $k \equiv 0.64$, which is the empirical proportion from the data. We fix the values of p_{in} and η_{tel}, but vary the values of p_{tel} and γ to study the effect of different telemedicine no-show versus abandonment rates on the optimal sequencing rule.

We first set $p_{in} = 20\%$ and $\eta_{tel} = 20\%$, i.e., the two modalities of visits have similar service incompletion rates as estimated in our partner clinics. In Figure 6, we consider low, median, and high abandonment rates scenarios. Our goal is to demonstrate which is the best policy for different values of p_{tel} and γ (shown in the top row of Figure 6). In the subsequent rows, we observe the degree of suboptimality when deriving the optimal sequencing rule based on the incorrect assumption of the degree of abandonment. In particular, in Figure 6, each row, denoted I, II, and III, corresponds to a specification of system dynamics for p_{tel} and γ. Each column, denoted a, b, and c, represents the degree of suboptimality in % incurred when applying the optimal sequencing derived based on specific system dynamics which could be different from the true system dynamics. The policies applied in each column correspond to the policy depicted in the first row of the figure. For example, in III(a), we apply the optimal policy derived for a system with $p_{tel} = 19\%$ and $\gamma = 3\%$ to a system with $p_{tel} = 1\%$ and $\gamma = 53\%$. In plots I(a), II(b), and III(c), we apply the correct optimal policy to the accurate system dynamics and thus have zero optimality gap. We make a few observations from the Figure. First when the abandonment cost dominates, Block Telemedicine First performs the best, regardless of the value of γ. When overtime cost dominates, the Block In-Person First performs the best, regardless of γ. When the waiting cost is large, the best-performing policy may depend on the abandonment rate. Applying the policy derived from a wrongly specified system can lead to highly suboptimal performance, i.e., 9% in plot III(a).

When the waiting cost is relatively large, applying the alternating policy can effectively utilize abandonment to help reduce intra-day delay. Figure 7 shows the performance of the

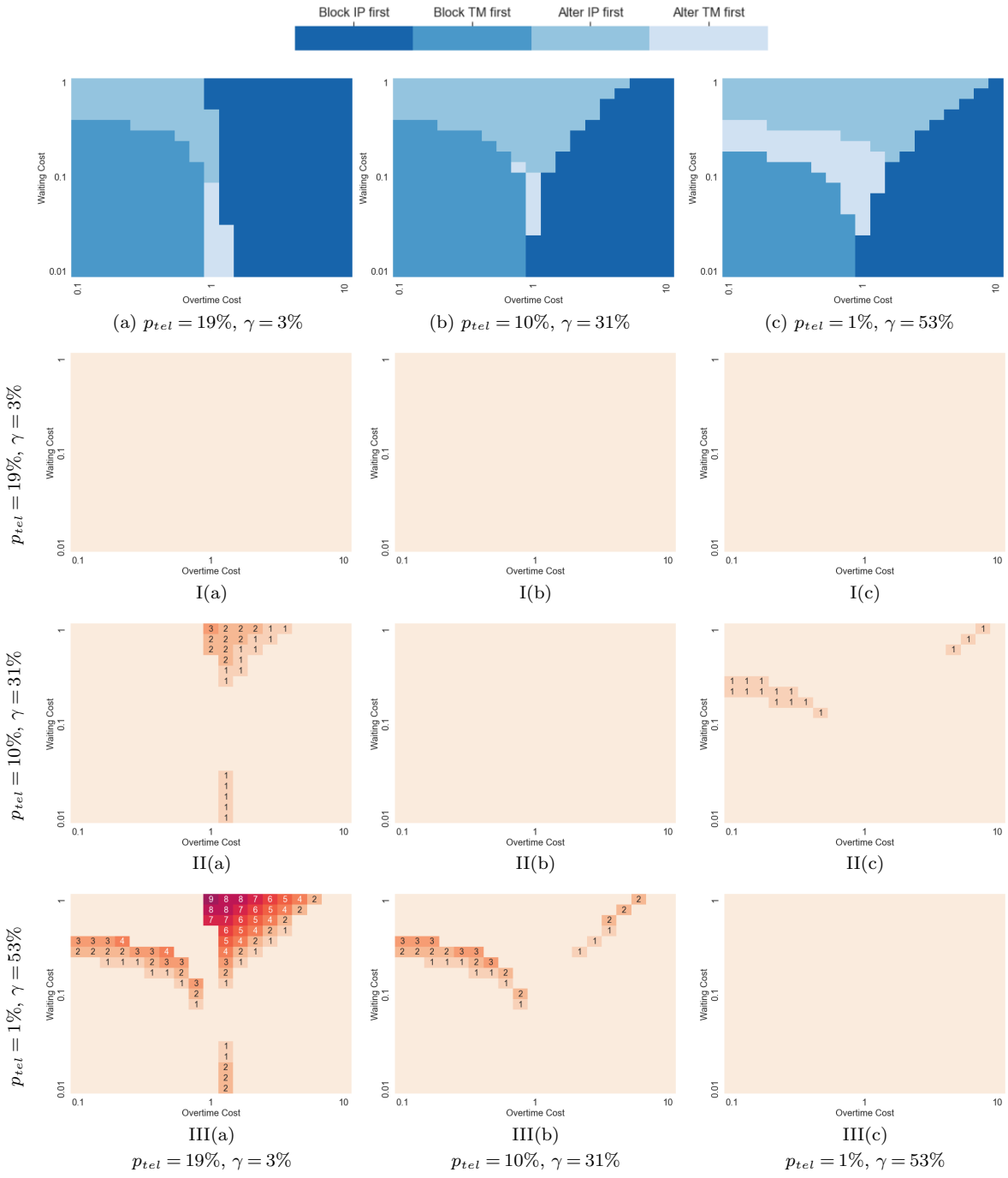
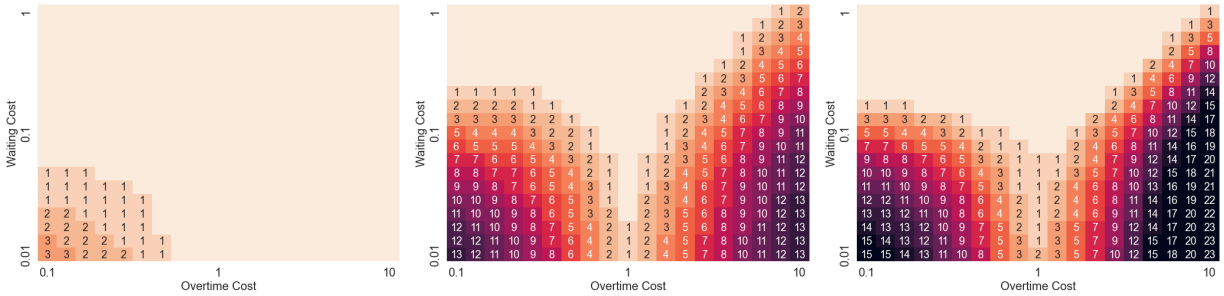


Figure 6 **First panel: best-performing heuristic sequencing rules. Second panel: the optimality gap in % when applying the policies derived based on the wrong system dynamics: each row of figures represents a specific system dynamic (under particular values of p_{tel} and γ), each column of figures represents the applying the policy depicted in the first panel to the system.** $p_{in} = 20\%$. $\eta_{tel} = 20\%$. $T = 30$. $N = 36$.

Alternating In-Person First policy, i.e, the optimality gap in % of applying Alternating In-Person First. We note that when the waiting cost is relatively large, i.e., $c_w > 0.1$ and $c_o < 1$,

Alternating In-Person First achieves very good performance. However, when the other costs dominate, Alternating In-Person First can lead to highly suboptimal performance, especially when the abandonment rate is high.



(a) $p_{tel} = 19\%$, $\gamma = 3\%$ (b) $p_{tel} = 10\%$, $\gamma = 31\%$ (c) $p_{tel} = 1\%$, $\gamma = 53\%$

Figure 7 **The optimality gap in % when applying the Alternating In-Person First policy.** $p_{in} = 20\%$. $\eta_{tel} = 20\%$. $T = 30$. $N = 36$.

We next consider a scenario where $p_{in} = 30\%$ and $\eta_{tel} = 20\%$, i.e., the in-person service incompletion rate (caused by no-show only) is much higher than the telemedicine incompletion rate. Figure 8 plots the best policy for different values of p_{tel} and γ (the first panel), and optimality gap in % when applying the policy derived based on the wrong system parameters (the second panel). We observe that the best-performing heuristic sequencing rules can be very different for different values of p_{tel} and γ. When the abandonment rate is low, incomplete visits are mostly due to no-shows. If overtime cost dominates the waiting cost, we should schedule the modality with a lower no-show rate first, i.e., Block Telemedicine First performs the best. If the waiting cost is high, we should schedule the modality with a higher no-show rate first. When the abandonment rate is high, if the overtime cost dominates, Block In-Person First performs the best; if the abandonment cost dominates, Block Telemedicine First performs the best. We also observe that not being able to correctly quantify the no-show versus abandonment rates can lead to highly suboptimal performance. The optimality gap can be as high as 35% (see plot III(a)). We also analyze the performance of the alternating rules in the three systems studied in Figure 8 (see Figure 17 in Appendix E). Similar to the case where p_{in} and η_{tel} are equal, when the waiting cost is relatively high, the alternating policy achieves robust and good performance.

We tested many different scenarios, i.e., with different values of p_{in} and η_{tel}, and different values of N, and the observations are similar to those in Figures 6, 7 and 8 (see more

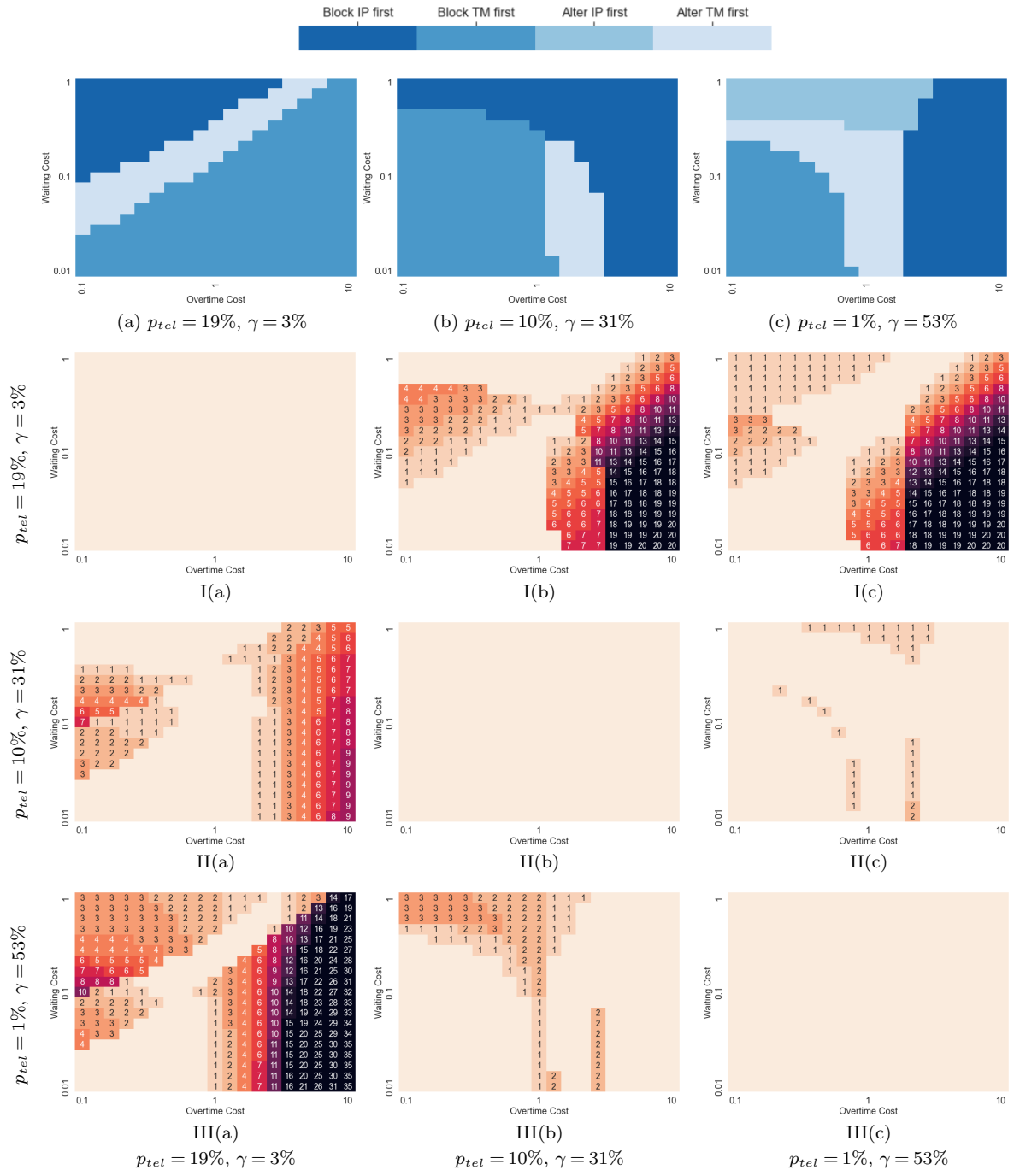


Figure 8 **First panel: best-performing heuristic sequencing policies. Second panel: the optimality gap in % when applying the policies derived based on the wrong system dynamics: each row of figures represents a specific system dynamic (under particular values of p_{tel} and γ), each column of figures represents the applying the policy depicted in the first panel to the system. $p_{in} = 30\%$. $\eta_{tel} = 20\%$. $T = 30$. $N = 36$.**

numerical experiments in Appendix E). Overall, when p_{in} and η_{tel} are similar and the waiting cost is low, we should apply Block Telemedicine First if the abandonment cost

dominates, and Block In-person First if the overtime cost dominates. When p_{in} and η_{tel} are very different, we need to accurately quantify the value of p_{tel} and γ to derive the optimal sequencing policy. Wrongly specifying the system parameters can lead to a large optimality gap. In general, when the waiting cost is relatively high, the alternating policies lead to robust and good performance.

6. Conclusions

In this paper, we examine the effect of physician availability on service incompletion for in-person and telemedicine appointments. Our findings reveal that intra-day delays significantly increase the likelihood of visit incompletion for telemedicine patients, but do not have a significant impact on the in-person visit incompletion rate. We attribute this increase in telemedicine visit incompletion to patients who leave without being seen and find that the estimated abandonment rate is 9.0%. After taking out the effect of abandonment, telemedicine patients have a lower no-show rate than in-person patients. To understand how different no-show and abandonment behaviors of telemedicine versus in-person patients affect how these patients should be sequenced within a day, we carry out a series of counterfactual experiments. Our analysis shows that it is important to take the abandonment behavior into consideration when making sequencing decisions. When the two modalities of visits have similar incompletion rates, the optimal sequencing rules remain largely consistent across different compositions of abandonment and no-show rates. However, when the two modalities of visits have very different incompletion rates, the optimal sequencing rules can be very different for different compositions of abandonment and no-show rates. In these scenarios, precise quantification of the abandonment rates is critical to avoid highly suboptimal sequencing decisions. The insights from the healthcare setting could potentially be extrapolated to other service settings, and could be particularly valuable in service systems where the providers want to integrate online services with in-person offerings.

For the estimation, we utilize a multivariate probit model with properly constructed instrumental variables to handle estimation challenges due to i) endogeneity, ii) missing values, and iii) one-sided misreporting of physician availability. To the best of our knowledge, our work is the first to handle all three estimation challenges simultaneously. These three estimation challenges may arise in many other observational studies, and our approach can be applied more broadly.

Our study has several limitations, which suggests interesting future research directions. First, our dataset is from outpatient clinics belonging to a single urban academic medical center. It would be interesting to look at other types of hospitals, which potentially cover different patient populations. Second, we capture the abandonment behavior indirectly by quantifying the impact of physician availability on the visit incompletion rate. This is mainly due to our data limitation. With more detailed and accurate data, it would be valuable to develop a better understanding of patients' abandonment behavior when waiting in person versus online. For example, we can characterize the patients' patience time distribution and study the mechanism behind the abandonment behavior. Third, patient no-show and abandonment have many important implications for appointment scheduling. In this work, we look at a specific aspect of appointment scheduling – intra-day sequencing and focus on heuristic sequencing rules only. It would be interesting and important to look into analytical or computational frameworks to optimally integrate telemedicine with in-person visits when considering both no-shows and abandonment, and the heterogeneous behaviors of the two modalities of visits.

References

Akşin Z, Ata B, Emadi SM, Su CL (2013) Structural estimation of callers' delay sensitivity in call centers. *Management Science* 59(12):2727–2746.

Alkilany R, Tarabichi Y, Hong R (2022) Telemedicine visits during covid-19 improved clinic show rates. *ACR open rheumatology* 4(2):136–141.

Allon G, Bassamboo A, Gurvich I (2011) "we will be right with you": Managing customer expectations with vague promises and cheap talk. *Operations research* 59(6):1382–1394.

Ansari S, Debo L, Ibanez M, Iravani S, Malik M, et al. (2022) Under-promising and over-delivering to improve patient satisfaction at emergency departments: Evidence from a field experiment providing wait information. *Available at SSRN 4135705* .

Arendt KW, Sadosty AT, Weaver AL, Brent CR, Boie ET (2003) The left-without-being-seen patients: what would keep them from leaving? *Annals of emergency medicine* 42(3):317–IN2.

Atar R, Giat C, Shimkin N (2010) The cμ/θ rule for many-server queues with abandonment. *Operations Research* 58(5):1427–1439.

Baiocchi M, Cheng J, Small DS (2014) Instrumental variable methods for causal inference. *Statistics in medicine* 33(13):2297–2340.

Batt RJ, Terwiesch C (2015) Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Science* 61(1):39–59.

Bayram A, Deo S, Iravani S, Smilowitz K (2020) Managing virtual appointments in chronic care. *IISE Transactions on Healthcare Systems Engineering* 10(1):1–17.

Berg BP, Murr M, Chermak D, Woodall J, Pignone M, Sandler RS, Denton BT (2013) Estimating the cost of no-shows and evaluating the effects of mitigation strategies. *Medical Decision Making* 33(8):976–985.

Bolandifar E, DeHoratius N, Olsen T, Wiler J (2019) An empirical study of the behavior of patients who leave the emergency department without being seen. *Journal of Operations Management* 65(5):430–446.

Bramati PS, Amaram-Davila JS, Reddy AS, Bruera E (2022) Reduction of missed palliative care appointments after the implementation of telemedicine. *Journal of pain and symptom management* .

Buell RW, Kim T, Tsay CJ (2017) Creating reciprocal value through operational transparency. *Management Science* 63(6):1673–1695.

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

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

Chan CW, Farias VF, Escobar GJ (2017) The impact of delays on service times in the intensive care unit. *Management Science* 63(7):2049–2072.

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

Delana K, Deo S, Ramdas K, Subburaman GBB, Ravilla T (2022) Multichannel delivery in healthcare: the impact of telemedicine centers in southern india. *Management Science* .

Franciosi EB, Tan AJ, Kassamali B, Leonard N, Zhou G, Krueger S, Rashighi M, LaChance A (2021) The impact of telehealth implementation on underserved populations and no-show rates by medical specialty during the covid-19 pandemic. *Telemedicine and e-Health* 27(8):874–880.

Friedman AB, Gervasi S, Song H, Bond AM, Chen AT, Bergman A, David G, Bailey JM, Brooks R, Smith-McLallen A (2022) Telemedicine catches on: changes in the utilization of telemedicine services during the covid-19 pandemic. *The American journal of managed care* 28(1):e1–e6.

Gallino S, Karacaoglu N, Moreno A (2022) Need for speed: The impact of in-process delays on customer behavior in online retail. *Operations Research* .

Greene WH (2003) *Econometric analysis* (Pearson Education India).

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

Hassin R, Mendel S (2008) Scheduling arrivals to queues: A single-server model with no-shows. *Management science* 54(3):565–572.

Hathaway BA, Emadi SM, Deshpande V (2022) Personalized priority policies in call centers using past customer interaction information. *Management Science* 68(4):2806–2823.

Heckman JJ (1979) Sample selection bias as a specification error. *Econometrica: Journal of the econometric society* 153–161.

Hu K, Xu X, Qu A (2021) The psychology of virtual queue: When waiting feels less like waiting. *Available at SSRN 3933040* .

Hui MK, Tse DK (1996) What to tell consumers in waits of different lengths: An integrative model of service evaluation. *Journal of Marketing* 60(2):81–90.

Karimi M, Lee EC, Couture SJ, Gonzales A, Grigorescu V, Smith SR, De Lew N, Sommers BD (2022) National survey trends in telehealth use in 2021: disparities in utilization and audio vs. video services.

Kong Q, Li S, Liu N, Teo CP, Yan Z (2019) Appointment scheduling under time-dependent patient no-show behavior. *Management Science (Forthcoming), Baruch College Zicklin School of Business Research Paper* (2019-04):03.

Lacy NL, Paulman A, Reuter MD, Lovejoy B (2004) Why we don't come: patient perceptions on no-shows. *The Annals of Family Medicine* 2(6):541–545.

Lee Y, Chen AN, Ilie V (2012) Can online wait be managed? the effect of filler interfaces and presentation modes on perceived waiting time online. *Mis Quarterly* 365–394.

Liu J, Xie J, Yang KK, Zheng Z (2019) Effects of rescheduling on patient no-show behavior in outpatient clinics. *Manufacturing & Service Operations Management* 21(4):780–797.

Liu N, Finkelstein SR, Kruk ME, Rosenthal D (2018) When waiting to see a doctor is less irritating: Understanding patient preferences and choice behavior in appointment scheduling. *Management Science* 64(5):1975–1996.

Liu N, Ziya S (2014) Panel size and overbooking decisions for appointment-based services under patient no-shows. *Production and Operations Management* 23(12):2209–2223.

Lokken TG, Blegen RN, Hoff MD, Demaerschalk BM (2020) Overview for implementation of telemedicine services in a large integrated multispecialty health care system. *Telemedicine and e-Health* 26(4):382–387.

Mandelbaum A, Zeltyn S (2013) Data-stories about (im) patient customers in tele-queues. *Queueing Systems* 75(2):115–146.

McKinsey & Company (2015) The virtual financial advisor: Delivering personalized advice in the digital age. URL `https://www.mckinsey.com/industries/financial-services/our-insights/the-virtual-financial-advisor-delivering-personalized-advice-in-the-digital-age#/`, accessed on June 20, 2023.

Moore CG, Wilson-Witherspoon P, Probst JC (2001) Time and money: effects of no-shows at a family practice residency clinic. *Family Medicine-Kansas City-* 33(7):522–527.

Newey WK, McFadden D (1994) Large sample estimation and hypothesis testing. *Handbook of econometrics* 4:2111–2245.

Nguimkeu P, Denteh A, Tchernis R (2019) On the estimation of treatment effects with endogenous misreporting. *Journal of econometrics* 208(2):487–506.

Osadchiy N, Kc D (2017) Are patients patient? the role of time to appointment in patient flow. *Production and Operations Management* 26(3):469–490.

Paquette S, Lin JC (2019) Outpatient telemedicine program in vascular surgery reduces patient travel time, cost, and environmental pollutant emissions. *Annals of vascular surgery* 59:167–172.

Park E, Ouyang H, Wang J, Savin S, Leung SC, Rainer T (2022) Patient sensitivity to emergency department waiting time announcements. *Available at SSRN 4047005* .

Patel SY, Mehrotra A, Huskamp HA, Uscher-Pines L, Ganguli I, Barnett ML (2021) Trends in outpatient care delivery and telemedicine during the covid-19 pandemic in the us. *JAMA internal medicine* 181(3):388–391.

Prékopa A (2013) *Stochastic programming*, volume 324 (Springer Science & Business Media).

Rajan B, Tezcan T, Seidmann A (2019) Service systems with heterogeneous customers: Investigating the effect of telemedicine on chronic care. *Management Science* 65(3):1236–1267.

Rozzi GC (2021) zipcoder: Advancing the analysis of spatial data at the zip code level in r. *Software impacts* 9:100099.

Russo JE, McCool RR, Davies L (2016) Va telemedicine: an analysis of cost and time savings. *Telemedicine and e-Health* 22(3):209–215.

Samorani M, Harris SL, Blount LG, Lu H, Santoro MA (2022) Overbooked and overlooked: machine learning and racial bias in medical appointment scheduling. *Manufacturing & Service Operations Management* 24(6):2825–2842.

Satiani B, Miller S, Patel D (2009) No-show rates in the vascular laboratory: analysis and possible solutions. *Journal of Vascular and Interventional Radiology* 20(1):87–91.

Schectman JM, Schorling JB, Voss JD (2008) Appointment adherence and disparities in outcomes among patients with diabetes. *Journal of general internal medicine* 23(10):1685–1687.

Snoswell CL, Comans TA (2021) Does the choice between a telehealth and an in-person appointment change patient attendance? *Telemedicine and e-Health* 27(7):733–738.

Statista (2021) Online legal services market size in the united states from 2011 to 2021 with forecast for 2022. URL `https://www.statista.com/statistics/1186899/online-legal-services-market-size-usa/`, accessed on June 20, 2023.

Sun S, Lu SF, Rui H (2020) Does telemedicine reduce emergency room congestion? evidence from new york state. *Information Systems Research* 31(3):972–986.

Sweis BM, Abram SV, Schmidt BJ, Seeland KD, MacDonald III AW, Thomas MJ, Redish AD (2018) Sensitivity to "sunk costs" in mice, rats, and humans. *Science* 361(6398):178–181.

Ülkü S, Hydock C, Cui S (2020) Making the wait worthwhile: Experiments on the effect of queueing on consumption. *Management Science* 66(3):1149–1171.

Van de Ven WP, Van Praag BM (1981) The demand for deductibles in private health insurance: A probit model with sample selection. *Journal of econometrics* 17(2):229–252.

Van der Vaart AW (2000) *Asymptotic statistics*, volume 3 (Cambridge university press).

Wooldridge JM (2010) *Econometric analysis of cross section and panel data* (MIT press).

Wootton R, Bahaadinbeigy K, Hailey D (2011) Estimating travel reduction associated with the use of telemedicine by patients and healthcare professionals: proposal for quantitative synthesis in a systematic review. *BMC health services research* 11(1):1–10.

World Economic Forum, Wood J (2022) World economic forum: These 3 charts show the global growth in online learning. URL `https://www.weforum.org/agenda/2022/01/online-learning-courses-reskill-skills-gap/`, accessed on June 20, 2023.

Yom-Tov GB, Yedidsion L, Xie Y (2021) An invitation control policy for proactive service systems: Balancing efficiency, value, and service level. *Manufacturing & Service Operations Management* 23(5):1077–1095.

Yu Q, Allon G, Bassamboo A (2017) How do delay announcements shape customer behavior? an empirical study. *Management Science* 63(1):1–20.

Yu Q, Allon G, Bassamboo A, Iravani S (2018) Managing customer expectations and priorities in service systems. *Management Science* 64(8):3942–3970.

Yu Q, Zhang Y, Zhou YP (2022) Delay information in virtual queues: A large-scale field experiment on a major ride-sharing platform. *Management Science* .

Zacharias C, Armony M (2017) Joint panel sizing and appointment scheduling in outpatient care. *Management Science* 63(11):3978–3997.

Zacharias C, Liu N, Begen MA (2022) Dynamic interday and intraday scheduling. *Operations Research* .

Zacharias C, Pinedo M (2014) Appointment scheduling with no-shows and overbooking. *Production and Operations Management* 23(5):788–801.

Zhang R, Han X, Wang R, Zhang J, Zhang Y (2022) Please don't make me wait! influence of customers' waiting preference and no-show behavior on appointment systems. *Production and Operations Management* .

Appendix

A. Variable Definitions and Summary Statistics

We control for a comprehensive list of patient-level and patient-visit-level variables in our models as discussed in Section 2.4. In this section, we provide detailed definitions of the variables in Table 5. We also present summary statistics of these variables in Table 6. In the case of continuous variables, we provide the means and standard deviations (in parentheses), while for discrete variables, we provide the number of observations and proportions (in parentheses) of each category.

Table 5: Description of some variables used in this study

Variable Name	Definition and Source
Partner	The variable is derived from the original variable "marital status." Specifically, the value is set to "yes" if the individual's marital status is listed as "married," "domestic partner," or "significant other." Conversely, the value is set to "no" if the individual's marital status is listed as "single," "divorced," "legally separated," or "widowed." In instances where the individual's marital status is listed as "other" or "unknown," the value of "partner" is recorded as "unknown."
Distance from home to clinic	The variable is calculated as the distance between the patient's home address and clinic address, as determined by Google Maps. If an accurate distance is not available, we construct the variable as the distance between the patient's home address Zip Code and the clinic Zip Code using the zip_distance function from the zipcodeR package.

Table 5: Description of some variables used in this study

Variable Name	Definition and Source
Median household income	The variable is derived from the patient's home address Zip Code. Specifically, the median household income within the Zip Code is obtained from the zip_code_db dataset within the "zipcodeR" package. This dataset is sourced from data.census.gov.
Time of day	The variable is derived from the appointment schedule. If the appointment is scheduled between 8 a.m. and 12 p.m., it is categorized as "morning." Conversely, if the appointment is scheduled between 12 p.m. (inclusive) and 6 p.m., it is categorized as "afternoon."
Lockdown	The variable is derived from the appointment date. If the appointment is scheduled before March 22, it is categorized as "pre-lockdown." If the appointment date is between March 22 and June 7, it is categorized as "lockdown." If the appointment date is between June 8 and July 19, it is categorized as "partially open." Finally, if the appointment date is after July 19, it is categorized as "open."

Table 6: Summary statistics of variables used in this study

Variable Name	In-Person (n=19,893)	Telemedicine (n=31,444)
Appointment Status		
Complete	15,863 (80%)	25,035 (80%)
Incomplete	4,030 (20%)	6,409 (20%)
Age, mean(SD)	61 (17)	61 (17)
Gender		
Male	6,102 (31%)	8,681 (28%)
Female	13,791 (69%)	22,763 (72%)
Race		
White, non-Hispanic	1,092 (5.5%)	1,634 (5.2%)
Black, non-Hispanic	2,179 (11%)	3,340 (11%)
Hispanic	12,902 (65%)	20,941 (67%)
AAPI	152 (0.8%)	267 (0.8%)
Others	1,283 (6.4%)	1,824 (5.8%)
Unknown	2,285 (11%)	3,438 (11%)
Insurance		
Commercial	1,957 (9.8%)	2,878 (9.2%)
Medicare	9,546 (48%)	15,106 (48%)
Medicaid	8,390 (42%)	13,460 (43%)
Partner		
No	14,122 (71%)	22,356 (71%)
Yes	5,500 (28%)	8,724 (28%)
Unknown	271 (1.4%)	364 (1.2%)
Distance (miles), mean (SD)	2.8 (3.5)	2.9 (3.5)
Income (thousands, annually), mean (SD)	38 (13)	38 (13)
Appointment Type		
New Patient	3,297 (17%)	893 (2.8%)
Follow Up	13,486 (68%)	30,551 (97%)
Hospital	354 (1.8%)	0 (0%)
Nurse	2,208 (11%)	0 (0%)
Same Day Urgent	323 (1.6%)	0 (0%)
Same Day Non-Urguent	225 (1.1%)	0 (0%)

Table 6: Summary statistics of variables used in this study

Variable Name	In-Person (n=19,893)	Telemedicine (n=31,444)
Lead Days, mean (SD)	17 (22)	9 (16)
Weekday		
Monday	4,363 (22%)	6,443 (20%)
Tuesday	4,047 (20%)	6,699 (21%)
Wednesday	3,753 (19%)	5,761 (18%)
Thursday	3,907 (20%)	6,687 (21%)
Friday	3,823 (19%)	5,854 (19%)
Time of Day		
Morning	10,952 (55%)	17,161 (55%)
Afternoon	8,941 (45%)	14,283 (45%)
Rush Hours		
No Rush	14,923 (75%)	23,038 (73%)
Morning Rush	4,311 (22%)	6,976 (22%)
Evening Rush	659 (3.3%)	1,430 (4.5%)
Lockdown		
Pre-Lockdown (pre-Mar22)	5,734 (29%)	428 (1.4%)
Lockdown (Mar22-Jun7)	602 (3.0%)	9,175 (29%)
Partially Open (Jun8-Jul18)	1,115 (5.6%)	6,342 (20%)
Open (post-July19)	12,442 (63%)	15,499 (49%)
Precipitation (inch per hour), mean (SD)	0.006 (0.033)	0.006 (0.036)
Snow Depth (inch), mean (SD)	0.148 (0.970)	0.134 (0.975)

B. Validity of Instrumental Variables

As discussed in Section 3.1.1, a valid instrumental variable needs to satisfy the relevance condition and the exclusion restriction. The exclusion restriction might be violated if there are unobserved confounders that are correlated with both the IV and the outcome variable. The two main potential confounders are patient severity and no-show history.

In this section, we run additional analyses focused on visits between August and December 2020 (the last five months of the dataset) in order to utilize the earlier months to construct various measures of patient risk. In the first analysis, we use the number of past medical visits as a proxy for patient severity. In particular, we calculate the number of visits within a 6-month window before the focal visit, which we refer to as $VisitHistory$. We run a linear regression with the IV $PreWorkProvider$ being the dependent variable, $VisitHistory$ being the treatment variable, and control for other patient-level and patient-visit-level information. The results are summarized in Table 7. We note that the coefficient for $VisitHistory$ is not significantly different from zero, which suggests that patient severity is unlikely to have a large confounding effect.

In the second analysis, we use a sliding window of 6 months to compute the patient's no-show history. We consider three no-show history measures: 1) No-show count (Count), which counts the total number of no-shows in the past 6 months; 2) No-show rate (Rate), which calculates the no-show rate over the past 6 months. If the patient does not have any appointments with the clinics over the past 6 months, we set the no-show count or rate to zero. 3) Bayesian no-show rate (Bayesian), where the prior follows a beta distribution with $\alpha = 3$ and $\beta = 12$ and we use the mean of the posterior distribution as a measure of the no-show history. Note that Beta$(3, 12)$ has a mean of 0.2 and a variance of 0.1. The mean matches the empirical mean we

Table 7 Estimates of the impact of the number of prior visits (as a proxy for severity) on instrumental variable (number of scheduled visits of the same provider prior to the focal visit). We report the estimated coefficients and the corresponding standard errors (in parentheses)

	In-Person	Telemedicine
$VisitHistory$	0.009	−0.011
	(0.008)	(0.007)
N	19893	31444

Note: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

observed in the data. For a patient who has k no-shows out of n appointments over the past 6 months, the posterior belief of no-show rate follows a beta distribution with $\alpha=3+k$ and $\beta=n-k+12$. For example, suppose there is one appointment for the focal patient over the past 6 months. If the patient showed up for that appointment, then the posterior mean would be 0.1875; if not, the posterior mean would be 0.25. We then regress the IV $PreWorkProvider$ on different measures of no-show history while controlling for other patient-level and patient-visit-level characteristics. Table 8 summarizes the estimation results. We note that in all cases, the coefficient for the no-show history is not significant, which suggests that the no-show history is also unlikely to have a large confounding effect.

Table 8 Estimates of the impact of no-show history on instrumental variable (number of scheduled visits of the same provider before the focal visit). We define no-show history in three different ways. We report the estimated coefficients and the corresponding standard errors (in parentheses)

	Count		Rate		Bayesian	
	In-Person	Telemedicine	In-Person	Telemedicine	In-Person	Telemedicine
$NoShowHistory$	−0.001	−0.009	−0.050	0.010	−0.183	−0.103
	(0.021)	(0.020)	(0.054)	(0.048)	(0.443)	(0.409)
N	19893	31444	19893	31444	19893	31444

Note: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

Lastly, as mentioned in Section 3.1.2, to ensure estimation stability, it is desirable to have a variable that is correlated with the selection probability but uncorrelated with visit incompletion (conditional on "selected") in the Heckman selection model. We choose the number of scheduled visits of the same clinic excluding the focal provider in a three-hour window around (1.5 hours before and 1.5 hours after) the focal visit, i.e., $WorkClinic$. The estimation results in Table 2 indicate that $WorkClinic$ is negatively correlated with $Observe$. We next run a logistic regression for $Incomplete$ on $WorkClinic$, while controlling for other patient-level and patient-visit-level covariates. The results are summarized in Table 9. We observe that $WorkClinic$ does not have a significant impact on $Incomplete$.

C. Additional Robustness Check

Table 10 summarizes the estimation results for some alternative specifications of physician availability discussed in Section 4.2. In particular, we define the physician to be unavailable if there is still a patient under their care 5/10/15 minutes after the scheduled start time of the focal visit.

Table 9 Estimates of the impact of the number of scheduled visits of the same clinic excluding the focal provider in a three-hour window around (1.5 hours before and 1.5 hours after) the focal visit on the likelihood of having an incomplete focal visit. We report the estimated coefficients and the corresponding standard errors (in parentheses)

	In-Person	Telemedicine
$WorkClinic$	0.0003	0.0001
	(0.0002)	(0.0001)
N	19893	31444

Note: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

Table 10 Full model estimates of the impact of physician availability using alternative definitions of physician availability. +5/10/15 min refers to defining a physician as being unavailable if there is still a patient under their care 5/10/15 minutes after the scheduled start time of the focal visit. We report the estimated average marginal effect and the corresponding standard error (in parentheses)

	+5min		+10min		+15min	
	In-Person	Telemedicine	In-Person	Telemedicine	In-Person	Telemedicine
Incomplete						
$Available$	0.019	-0.073^{***}	0.020	-0.071^{***}	0.024	-0.070^{***}
	(0.019)	(0.020)	(0.020)	(0.019)	(0.020)	(0.022)
Available						
$PreWorkProvider$	-0.079^{***}	-0.081^{***}	-0.080^{***}	-0.074^{***}	-0.078^{***}	-0.068^{***}
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
ReportTrue						
$RelWorkClinic$	-0.499^{***}	-0.496^{***}	-0.466^{***}	-0.457^{***}	-0.441^{***}	-0.418^{***}
	(0.014)	(0.022)	(0.015)	(0.020)	(0.015)	(0.024)
Observe						
$WorkClinic$	-0.001^{***}	-0.002^{***}	-0.001^{***}	-0.002^{***}	-0.001^{***}	-0.002^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
ρ_{12}	0.126^{***}	0.193^{***}	0.095^{**}	0.172^{***}	0.111^{**}	0.150^{**}
ρ_{13}	-0.033	0.134	-0.051	0.087	-0.048	0.034
ρ_{14}	-0.004	0.099	-0.010	0.101^{*}	-0.036	0.094
N	19893	31444	19893	31444	19893	31444

Note: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$

D. Proof of Proposition 1

We first introduce some notations used in the proof. Let $\|\cdot\|_2$ denote the l_2 norm and $\|\cdot\|_\infty$ denote the l_∞ norm for vectors. Let $\|\cdot\|_F$ denote Frobenius norm for matrices. Denote $y_{10}=\beta_1^\top X_1$, $y_{11}=\beta_1^\top X_1+\delta$, $y_2=\beta_2^\top X_2+\alpha_2 PreWorkProvider$, $y_3=\beta_3^\top X_3+\alpha_3 RelWorkClinic$, and $y_4=\beta_4^\top X_4+\alpha_4 WorkClinic$. Recall that $AR=Available*ReportTrue$, $z=(X^\top, Incomplete, AR, Observe)^\top$ when $Observe=1$, $z=(X^\top, Incomplete, Observe)^\top$ when $Observe=0$, and $\theta=(\beta_1^\top,\delta,\beta_2^\top,\alpha_2,\beta_3^\top,\alpha_3,\beta_4^\top,\alpha_4,\rho_{12},\cdots,\rho_{34})^\top$. Then,

$$
\begin{aligned}
&f(z;\theta)\\
=&\Phi_4(y_{11},y_2,y_3,y_4;\Sigma_1)^{Incomplete\cdot AR\cdot Observe}\Phi_4(-y_{11},y_2,y_3,y_4;\Sigma_2)^{(1-Incomplete)\cdot AR\cdot Observe}\\
&[\Phi_4(y_{11},y_2,-y_3,y_4;\Sigma_3)+\Phi_4(y_{10},-y_2,y_3,y_4;\Sigma_4)+\Phi_4(y_{10},-y_2,-y_3,y_4;\Sigma_5)]^{Incomplete\cdot(1-AR)\cdot Observe}
\end{aligned}
$$

$[\Phi_4(-y_{11}, y_2, -y_3, y_4; \Sigma_6) + \Phi_4(-y_{10}, -y_2, y_3, y_4; \Sigma_7) + \Phi_4(-y_{10}, -y_2, -y_3, y_4; \Sigma_8)]^{(1-Incomplete)\cdot(1-AR)\cdot Observe}$
$\Phi(-y_4)^{(1-Observe)}$,

where Φ denotes the cumulative distribution function (CDF) for the univariate standard Gaussian, $\Phi_4(\cdot;\Sigma)$ denotes the CDF of a 4-dimensional Gaussian vector with mean zero and covariance matrix Σ, and

$$\Sigma_1 = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 \end{pmatrix}, \ \Sigma_2 = \begin{pmatrix} 1 & -\rho_{12} & -\rho_{13} & -\rho_{14} \\ -\rho_{12} & 1 & \rho_{23} & \rho_{24} \\ -\rho_{13} & \rho_{23} & 1 & \rho_{34} \\ -\rho_{14} & \rho_{24} & \rho_{34} & 1 \end{pmatrix},$$

$$\Sigma_3 = \begin{pmatrix} 1 & \rho_{12} & -\rho_{13} & \rho_{14} \\ \rho_{12} & 1 & -\rho_{23} & \rho_{24} \\ -\rho_{13} & -\rho_{23} & 1 & -\rho_{34} \\ \rho_{14} & \rho_{24} & -\rho_{34} & 1 \end{pmatrix}, \ \Sigma_4 = \begin{pmatrix} 1 & -\rho_{12} & \rho_{13} & \rho_{14} \\ -\rho_{12} & 1 & -\rho_{23} & -\rho_{24} \\ \rho_{13} & -\rho_{23} & 1 & \rho_{34} \\ \rho_{14} & -\rho_{24} & \rho_{34} & 1 \end{pmatrix}, \ \Sigma_5 = \begin{pmatrix} 1 & -\rho_{12} & -\rho_{13} & \rho_{14} \\ -\rho_{12} & 1 & \rho_{23} & -\rho_{24} \\ -\rho_{13} & \rho_{23} & 1 & -\rho_{34} \\ \rho_{14} & -\rho_{24} & -\rho_{34} & 1 \end{pmatrix},$$

$$\Sigma_6 = \begin{pmatrix} 1 & -\rho_{12} & \rho_{13} & -\rho_{14} \\ -\rho_{12} & 1 & -\rho_{23} & \rho_{24} \\ \rho_{13} & -\rho_{23} & 1 & -\rho_{34} \\ -\rho_{14} & \rho_{24} & -\rho_{34} & 1 \end{pmatrix}, \ \Sigma_7 = \begin{pmatrix} 1 & \rho_{12} & -\rho_{13} & -\rho_{14} \\ \rho_{12} & 1 & -\rho_{23} & -\rho_{24} \\ -\rho_{13} & -\rho_{23} & 1 & \rho_{34} \\ -\rho_{14} & -\rho_{24} & \rho_{34} & 1 \end{pmatrix}, \ \Sigma_8 = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & -\rho_{14} \\ \rho_{12} & 1 & \rho_{23} & -\rho_{24} \\ \rho_{13} & \rho_{23} & 1 & -\rho_{34} \\ -\rho_{14} & -\rho_{24} & -\rho_{34} & 1 \end{pmatrix}.$$

For the brevity of notation, the variance-covariance matrix in Φ_4 will be suppressed in the following proof. We also write $l(z;\theta) = \ln f(z;\theta)$. We first introduce two auxiliary lemmas. The first lemma shows that the model is identifiable.

LEMMA 1. *Given the assumptions in Proposition 1, for any $\theta \neq \theta_0$, $\mathbb{P}(f(z;\theta) \neq f(z;\theta_0)) > 0$.*

Proof of Lemma 1 For $\tilde{\theta} = (\tilde{\beta}_1^\top, \tilde{\delta}, \tilde{\beta}_2^\top, \tilde{\alpha}_2, \tilde{\beta}_3^\top, \tilde{\alpha}_3, \tilde{\beta}_4^\top, \tilde{\alpha}_4, \tilde{\rho}_{12}, \cdots, \tilde{\rho}_{34})^\top$, denote $\tilde{y}_{10} = \tilde{\beta}_1^\top X_1$, $\tilde{y}_{11} = \tilde{\beta}_1^\top X_1 + \tilde{\delta}$, $\tilde{y}_2 = \tilde{\beta}_2^\top X_2 + \tilde{\alpha}_2 PreWorkProvider$, $\tilde{y}_3 = \tilde{\beta}_3^\top X_3 + \tilde{\alpha}_3 RelWorkClinic$, and $\tilde{y}_4 = \tilde{\beta}_4^\top X_4 + \tilde{\alpha}_4 WorkClinic$. Let $y_1 = y_{10} + (y_{11} - y_{10}) \cdot AR$. Then, it suffices to show that for any $\tilde{\theta} \neq \theta$,

$$\mathbb{P}(\Phi_4(y_1, y_2, y_3, y_4) \neq \Phi_4(\tilde{y}_1, \tilde{y}_2, \tilde{y}_3, \tilde{y}_4)) > 0.$$

The nonsingularity of $\mathbb{E}[X_i X_i^\top]$ implies that $\beta^\top \mathbb{E}[X_i X_i^\top]\beta > 0$ for any $\beta \neq 0$. Thus, for $\beta_1 \neq \tilde{\beta}_1$,

$$\mathbb{E}[((\beta_1 - \tilde{\beta}_1)^\top X_{1i})^2] = (\beta_1 - \tilde{\beta}_1)^\top \mathbb{E}[X_{1i} X_{1i}^\top](\beta_1 - \tilde{\beta}_1) > 0,$$

which implies that $\mathbb{P}((\beta_1 - \tilde{\beta}_1)^\top X_{1i} \neq 0) > 0$, i.e.,

$$\mathbb{P}(\beta_1^\top X_{1i} \neq \tilde{\beta}_1^\top X_{1i}) > 0.$$

Then, for any $(\beta_1, \delta) \neq (\tilde{\beta}_1, \tilde{\delta})$,

$$\mathbb{P}(\Phi_4(y_1, y_2, y_3, y_4) \neq \Phi_4(\tilde{y}_1, y_2, y_3, y_4)) > 0.$$

Following similar lines of argument, we have when $(\beta_2, \alpha_2) \neq (\tilde{\beta}_2, \tilde{\alpha}_2)$,

$$\mathbb{P}(\Phi_4(y_1, y_2, y_3, y_4) \neq \Phi_4(y_1, \tilde{y}_2, y_3, y_4)) > 0.$$

When $(\beta_3, \alpha_3) \neq (\tilde{\beta}_3, \tilde{\alpha}_3)$,

$$\mathbb{P}(\Phi_4(y_1, y_2, y_3, y_4) \neq \Phi_4(y_1, y_2, \tilde{y}_3, y_4)) > 0.$$

When $(\beta_4, \alpha_4) \neq (\tilde{\beta}_4, \tilde{\alpha}_4)$,

$$\mathbb{P}(\Phi_4(y_1, y_2, y_3, y_4) \neq \Phi_4(y_1, y_2, y_3, \tilde{y}_4)) > 0.$$

Since y_1 and y_2 contain different covariates, for any $(\beta_1, \delta, \beta_2, \alpha_2) \neq (\tilde{\beta}_1, \tilde{\delta}, \tilde{\beta}_2, \tilde{\alpha}_2)$,

$$\mathbb{P}(\Phi_4(y_1, y_2, y_3, y_4) \neq \Phi_4(\tilde{y}_1, \tilde{y}_2, y_3, y_4)) > 0.$$

Similarly, since y_1, y_2, y_3, y_4 all contain different covariates, for any $\theta \neq \tilde{\theta}$,

$$\mathbb{P}(\Phi_4(y_1, y_2, y_3, y_4) \neq \Phi_4(\tilde{y}_1, \tilde{y}_2, \tilde{y}_3, \tilde{y}_4)) > 0.$$

(Note that the above argument is not valid if y_2 and y_3 contain exactly the same covariates, i.e., $y_2 = \beta_2^\top X_2 + \alpha_2 IV$ and $y_3 = \beta_3^\top X_3 + \alpha_3 IV$. In this case,

$$\Phi_4(y_1, \tilde{\beta}_3^\top X_2 + \tilde{\alpha}_3 IV, \tilde{\beta}_2^\top X_3 + \tilde{\alpha}_2 IV, y_4) = \Phi_4(y_1, \tilde{\beta}_2^\top X_2 + \tilde{\alpha}_2 IV, \tilde{\beta}_3^\top X_3 + \tilde{\alpha}_3 IV, y_4)$$

since X_2 and X_3 contain the same covariates.)

Lastly, changing any parameters in the variance-covariance matrix also changes the value of the Φ_4. Hence, the model is identifiable. □

LEMMA 2. *There exists a quadratic function of $\|X\|_2$, $d(X)$, such that $|l(z;\theta)| \leq d(X)$.*

Proof of Lemma 2 We first show that the last component of the log-likelihood function $|\ln \Phi(-y_4)|$ is dominated by a quadratic function of $\|X\|_2$. From Example 1.2 in Newey and McFadden (1994), the derivative of $\ln \Phi(\cdot)$, i.e.,

$$\frac{d \ln \Phi(u)}{du} = \frac{\phi(u)}{\Phi(u)},$$

is a convex function. In addition,

$$\lim_{u \to -\infty} \frac{\phi(u)}{\Phi(u)} + u = 0 \text{ and } \lim_{u \to +\infty} \frac{\phi(u)}{\Phi(u)} = 0.$$

Thus, $\frac{\phi(u)}{\Phi(u)} \leq c_1(1 + |u|)$ for some constant c_1. Since Φ is continuously differentiable, by the mean value theorem, there exists $\bar{y}_4$ with $|\bar{y}_4| \in (0, |y_4|)$, such that

$$\begin{aligned}
|\ln \Phi(-y_4)| &= \left| \ln \Phi(0) + \frac{\phi(-\bar{y}_4)}{\Phi(-\bar{y}_4)} y_4 \right| \\
&\leq |\ln \Phi(0)| + \frac{\phi(-\bar{y}_4)}{\Phi(-\bar{y}_4)} |y_4| \\
&\leq |\ln \Phi(0)| + c_1(1 + |\bar{y}_4|)|y_4| \\
&\leq |\ln \Phi(0)| + c_1(1 + |y_4|)|y_4| \\
&\leq |\ln \Phi(0)| + c_1 \left(1 + \|X\|_2 \left\| \begin{pmatrix} \beta_4 \\ \alpha_4 \end{pmatrix} \right\|_2 \right) \|X\|_2 \left\| \begin{pmatrix} \beta_4 \\ \alpha_4 \end{pmatrix} \right\|_2,
\end{aligned}$$

where the last inequality follows from Cauchy–Schwarz inequality. Since the parameter space is compact, $|\ln \Phi(-y_4)|$ is dominated by a polynomial function of $\|X\|_2$ up to degree 2.

Next, we show that the argument for the one-dimensional case can be extended to the multivariate case. We illustrate this with the bivariate case only, since the extension to higher dimensions follows the same lines of argument. First, we note that for any $y_1 = \beta_1^\top X_1$, $y_2 = \beta_2^\top X_2$, and $y = \min(y_1, y_2)$, we have $|\ln \Phi(y_1, y_2)| \leq |\ln \Phi(y, y)|$. Define $\tilde{\Phi}(u) = \Phi(u, u)$ and $\tilde{\phi}(u) = \phi(u, u)$. Similar to the univariate case, $\frac{\tilde{\phi}(u)}{\tilde{\Phi}(u)}$ is convex, $\lim_{u \to -\infty} \frac{\tilde{\phi}(u)}{\tilde{\Phi}(u)} - u = 0$, and $\lim_{x \to +\infty} \frac{\tilde{\phi}(u)}{\tilde{\Phi}(u)} = 0$. Thus, $|\log \tilde{\Phi}(y)|$ is dominated by a quadratic function of $|y|$. Since

$$|y| \leq \left\| \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \right\|_\infty = \left\| \begin{pmatrix} \beta_1^\top X_1 \\ \beta_2^\top X_2 \end{pmatrix} \right\|_\infty \leq \sqrt{(\beta_1^\top X_1)^2 + (\beta_2^\top X_2)^2} \leq \left\| \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \right\|_2 \left\| \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \right\|_2,$$

$|\ln \Phi(y_1, y_2)|$ is dominated by a polynomial function of $\|X\|_2$ up to degree 2. □

Proof of Proposition 1 Let $l_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ln f(z_i;\theta)$ and $l_0(\theta) = \mathbb{E}[\ln f(z_i;\theta)]$. The maximum likelihood estimator $\hat{\theta}_n$ is defined as

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta} l_n(\theta).$$

We first note that by the Weak Law of Large Numbers,

$$l_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ln f(z_i;\theta) \xrightarrow{p} \mathbb{E}[\ln f(z_i;\theta)] = l_0(\theta).$$

We also note that $l_0(\theta)$ is continuous because it is a linear combination of the logarithms of Normal CDFs.

To prove the consistency of $\hat{\theta}_n$, by Theorem 5.7 in Van der Vaart 2000, it suffices to show that (a) θ_0 is a well-separated point of maximum of $l_0(\theta)$, i.e., for any complement of open neighbourhood of θ_0, $\bar{\mathcal{N}_0}$,

$$\sup_{\theta\in\bar{\mathcal{N}_0}} l_0(\theta) < l_0(\theta_0);$$

and (b) $l_n(\theta)$ converges uniformly to $l_0(\theta)$ in probability for $\theta\in\Theta$, i.e.,

$$\sup_{\theta\in\Theta} |l_n(\theta) - l_0(\theta)| \xrightarrow{p} 0.$$

For a compact set Θ and a continuous function $l_0(\theta)$, condition (a) is equivalent to that θ_0 is the unique maximizer of $l_0(\theta)$. Lemma 1 shows that

$$\forall\theta\neq\theta_0 \Longrightarrow \mathbb{P}(f(z;\theta)\neq f(z;\theta_0)) > 0.$$

By Lemma 5.35 in Van der Vaart (2000), the identifiablity implies that $l_0(\theta)$ attains its maximum uniquely at θ_0. Thus, condition (a) holds.

Recall that $l(z;\theta) = \ln f(z;\theta)$. Condition (b) is equivalent to the set of functions $\{l(z;\theta):\theta\in\Theta\}$ being *Glivenko-Cantelli* (Van der Vaart 2000), and a set of sufficient conditions for being Glivenko-Cantelli is that (b1) Θ is compact; (b2) $l(z;\theta)$ is continuous in θ for any z; and (b3) $l(z;\theta)$ is dominated by an integrable function.

Condition (b1) is the assumption of Proposition 1. Condition (b2) holds because $l(z;\theta)$ is a linear combination of the logarithm of Normal CDFs as discussed above. Condition (b3) holds due to Lemma 2, i.e., $|l(z;\theta)|$ can be bounded by a quadratic function of $\|X\|_2$.

To prove the asymptotic normality of $\hat{\theta}_n$, by Theorem 3.3 in Newey and McFadden (1994), we only need to verify that (i) $f(z_i;\theta)$ is twice continuously differentiable and there exists an open neighborhood of θ_0, $\mathcal{N}_0$, such that $f(z_i;\theta)>0$ for all $\theta\in\mathcal{N}_0$; (ii) $\int \sup_{\theta\in\mathcal{N}_0}\|\nabla_\theta f(z_i;\theta)\|_2 dz < \infty$ and $\int \sup_{\theta\in\mathcal{N}_0}\|\nabla^2_\theta f(z_i;\theta)\|_F dz < \infty$. (iii) $\mathbb{E}[(\nabla_\theta \ln f(z_i;\theta_0))(\nabla_\theta \ln f(z_i;\theta_0))^\top]$ exists; (iv) $\mathbb{E}[\sup_{\theta\in\mathcal{N}_0}\|\nabla^2_\theta \ln f(z_i;\theta))\|_F] < \infty$. We next verify conditions (i) to (iv) one by one. We first re-write $f(z;\theta)$ as

$$\begin{aligned}
f(z;\theta) =& Incomplete\cdot AR\cdot Observe\cdot \Phi_4(y_{11},y_2,y_3,y_4)+\\
&(1-Incomplete)\cdot AR\cdot Observe\cdot \Phi_4(-y_{11},y_2,y_3,y_4)+\\
&Incomplete\cdot(1-AR)\cdot Observe\cdot\\
&[\Phi_4(y_{11},y_2,-y_3,y_4)+\Phi_4(y_{10},-y_2,y_3,y_4)+\Phi_4(y_{10},-y_2,-y_3,y_4)]+\\
&(1-Incomplete)\cdot(1-AR)\cdot Observe\cdot\\
&[\Phi_4(-y_{11},y_2,-y_3,y_4)+\Phi_4(-y_{10},-y_2,y_3,y_4)+\Phi_4(-y_{10},-y_2,-y_3,y_4)]+\\
&(1-Observe)\Phi(-y_4).
\end{aligned}$$

First, we note that the partial derivative of the CDF of a multivariate distribution is connected to its conditional CDF and PDF through the following form Prékopa (2013):

$$\frac{\partial}{\partial \xi_i}\text{CDF}(\xi_1, \xi_2, \cdots, \xi_n) = \text{PDF}(\xi_i)\text{CDF}(\xi_1, \xi_2, \cdots, \xi_{i-1}.\xi_{i+1}, \cdots, \xi_n|\xi_i).$$

In the case of normal distributions, the partial derivative of the CDF of a n-dimensional Normal is the product of the PDF of a univariate Normal and the CDF of an $(n-1)$-dimensional Normal. Normal PDFs and CDFs are bounded and continuously differentiable. For the brevity of notation, we use Φ_2 and Φ_3 to denote the CDFs of generic 2- and 3-dimensional zero-mean Normal random vectors respectively.

Condition (i) holds because $f(z_i;\theta)$ is a combination of Normal CDFs.

For condition (ii), we first note that $\nabla_\theta f(z_i;\theta)$ is a vector with each element being a linear combination of functions of the form $\phi \cdot \Phi_3$. For example,

$$\begin{aligned}\frac{\partial f(z_i;\theta)}{\partial \alpha_2} =& (Incomplete_i \cdot AR_i \cdot Observe_i) \cdot PreWorkProvider_i \cdot \phi \cdot \Phi_3 + \\ & (1 - Incomplete_i) \cdot AR_i \cdot Observe_i \cdot PreWorkProvider_i \cdot \phi \cdot \Phi_3 + \cdots\end{aligned}$$

which is a linear combination of $\phi \cdot \Phi_3$ weighted by elements of z_i. Since $Incomplete_i$, AR_i, $Observe_i$, ϕ, and Φ_3 are all bounded by 1,

$$\sup_{\theta \in \mathcal{N}_0} \left| \frac{\partial f(z_i;\theta)}{\partial \alpha_2} \right| \leq k_1 PreWorkProvider_i.$$

for some constant k_1. Thus,

$$\int \sup_{\theta \in \mathcal{N}_0} \|\nabla_\theta f(z_i;\theta)\|_2 dz \leq c_1 \mathbb{E}[\|X_i\|_2] < \infty$$

for some constant c_1. Similarly, we note that $\nabla^2_\theta f(z_i;\theta)$ is a matrix with each element being a linear combination of functions of the form $(\phi'\Phi_3 + \phi^2\Phi_2)$. For example,

$$\begin{aligned}\frac{\partial^2 f(z_i;\theta)}{\partial \alpha_2^2} =& (Incomplete_i \cdot AR_i \cdot Observe_i) \cdot PreWorkProvider_i^2 \cdot (\phi'\Phi_3 + \phi^2\Phi_2) + \\ & (1 - Incomplete_i) \cdot AR_i \cdot Observe_i \cdot PreWorkProvider_i^2 \cdot (\phi'\Phi_3 + \phi^2\Phi_2) + \cdots\end{aligned}$$

Since $Incomplete_i$, AR_i, $Observe_i$, ϕ, Φ_2, and Φ_3 are all bounded by some constant, we have

$$\sup_{\theta \in \mathcal{N}_0} \left| \frac{\partial^2 f(z_i;\theta)}{\partial \alpha_2^2} \right| \leq k_2 PreWorkProvider_i^2.$$

for some constant k_2 and

$$\int \sup_{\theta \in \mathcal{N}_0} \|\nabla^2_\theta f(z_i;\theta)\|_F dz \leq c_2 \mathbb{E}[\|X_i X_i^\top\|_F] < \infty$$

for some constant c_2. Thus, condition (ii) holds.

For condition (iii), we first note that

$$\nabla_\theta \ln f(z_i;\theta_0) = \frac{\nabla_\theta f(z_i;\theta_0)}{f(z_i;\theta_0)}.$$

We have shown, in verifying condition (ii), that each element in the numerator is bounded by a multiple of an element in z_i. The denominator is a linear combination of Normal CDFs and both z_i and θ are bounded. Hence, there exits $\epsilon > 0$ such that

$$\sup_{z_i, \theta \in \Theta} f(z_i;\theta) > \epsilon.$$

Then, each element in the matrix $(\nabla_\theta \ln f(z_i;\theta_0))(\nabla_\theta \ln f(z_i;\theta_0))^\top$ is bounded by a constant and condition (iii) holds.

For condition (iv), we first note that

$$\nabla^2_\theta \ln f(z_i;\theta) = \frac{\nabla^2_\theta f(z_i;\theta)}{f(z_i;\theta)} - \frac{\nabla_\theta f(z_i;\theta)\nabla_\theta f(z_i;\theta)^\top}{(f(z_i;\theta))^2}.$$

The numerators, as shown in the proof of condition (ii), are matrices with bounded elements. The denominators, as shown in the proof of condition (iii), are positive and bounded away from zero. Thus, $\sup_{\theta\in\mathcal{N}_0} \nabla^2_\theta \ln f(z_i;\theta)$ is bounded and condition (iv) holds. □

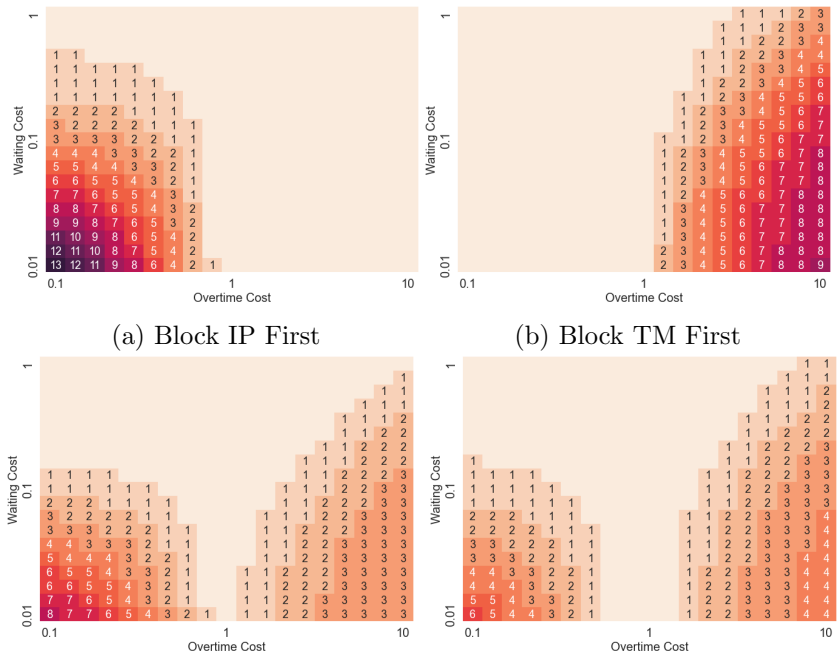
E. Additional Numerical Results for the Counterfactual Analyses

In this section, we provide additional counterfactual experiments to support the findings about appointment sequencing in Section 5.

First, in the setting of our partner clinics, the service incompletion rates for in-person and telemedicine patients are both around 20%. If all service incompletions are caused by no-shows, then there is no difference in performance among the four heuristic sequencing rules, i.e., they will lead to the same average cost. Figure 9 shows the optimality gap in % of different sequencing rules under the true system dynamics, i.e., the telemedicine patients can abandon when delayed. We note that there can be a substantial optimality gap (6% – 13%) under some cost parameters. We also note that alternating telemedicine first achieves fairly robust performance across different cost parameters. Note that when in-person has a relatively high no-show rate, by alternating between in-person and telemedicine visits, telemedicine patients are less likely to experience physician delays. On the other hand, if the physician is delayed, the abandonment behavior of telemedicine patients can help reduce physician overtime.

Second, recall that p_{in} denotes the in-person service no-show rate (which is the same as the in-person service incompletion rate), η_{tel} denotes the telemedicine service incompletion rate, p_{tel} denotes the telemedicine no-show rate, and γ denotes the telemedicine abandonment rate. In Figures 10 – 15, we follow a similar approach as that in Figures 6 and 8 to conduct some further sensitivity analysis. In particular, we fix the values of p_{in} and η_{tel} while varying p_{tel} and γ.

Recall that Figure 6 considers a clinic with $p_{in} = \eta_{tel} = 20\%$ and $N = 36$. Figure 10 maintains the same service incompletion rates but increases the panel size N to 42 to represent a more crowded (overbooked) clinic. Figure 11 keeps the same panel size, i.e., $N = 36$, but increases the service incompletion rates to $p_{in} = \eta_{tel} = 30\%$. Figure 12 sets $p_{in} = \eta_{tel} = 30\%$ and $N = 42$. In each figure, the first panel shows the best-performing heuristic sequencing rules for various abandonment scenarios. We observe that when p_{in} and η_{tel} are similar and the waiting cost is low, we should apply Block Telemedicine First if the abandonment cost dominates, and Block In-person First if the overtime cost dominates, regardless of γ. When the waiting cost is large, the best-performing policy may depend on the abandonment rate. The second panel shows the optimality gap when applying the policies derived based on the wrong system dynamics. There, each row represents a specific system dynamic, and each column represents applying the policy depicted in the first panel. We observe that applying the policy derived based on a wrongly specified system can lead to a large



(a) Block IP First (b) Block TM First

(c) Alternating IP First (d) Alternating TM First

Figure 9 **Optimality gap in % of each of the four heuristic sequencing rules.** $p_{in} = 20\%$, $p_{tel} = 18\%$, $\gamma = 9\%$, $T = 30$. $N = 36$.

optimality gap (8% in Figure 10 III(a), 21% in Figure 11 III(a), 29% in Figure 12 III(a)). These findings are consistent with the observations in Figure 6.

Recall that Figure 8 considers a clinic with $p_{in} = 30\%$, $\eta_{tel} = 20\%$ and $N = 36$. Figure 13 keeps the same service incompletion rates but increases the panel size to $N = 42$. Figure 14 keeps $N = 36$, but sets $p_{in} = 20\%$ and $\eta_{tel} = 10\%$. Figure 15 sets $p_{in} = 20\%$, $\eta_{tel} = 10\%$, and $N = 42$. We observe from the first panels that the best-performing heuristic sequencing rule varies with different p_{tel} and γ values. For low abandonment rates, Block Telemedicine First is preferred when the overtime cost dominates, while Block In-Person First is preferred when the waiting cost dominates. For high abandonment rates, Block In-Person First performs the best when the overtime cost dominates, while Block Telemedicine First performs the best when the abandonment cost dominates. In addition, when p_{in} and η_{tel} are very different, wrongly specifying the system parameters (p_{tel} and η) can lead to a very large optimality gap (44% in Figure 13 III(a), 13% in Figure 14 I(c), and 9% in Figure 15 III(a)). These findings are consistent with the observations in Figure 8.

Lastly, Figures 16 and 17 demonstrate the performance (optimality gap) of the alternating sequencing rules for different system parameters, i.e., different service incompletion rates, telemedicine no-show rates and abandonment rates. Similar to what we have observed in Figure 7, when the waiting cost is relatively high (i.e., $c_w > 0.1$ and $c_o < 1$), the worst-case optimality gaps of the alternating sequence rules are small, no larger than 4%, suggesting that these policies are robust in this parameter regime.

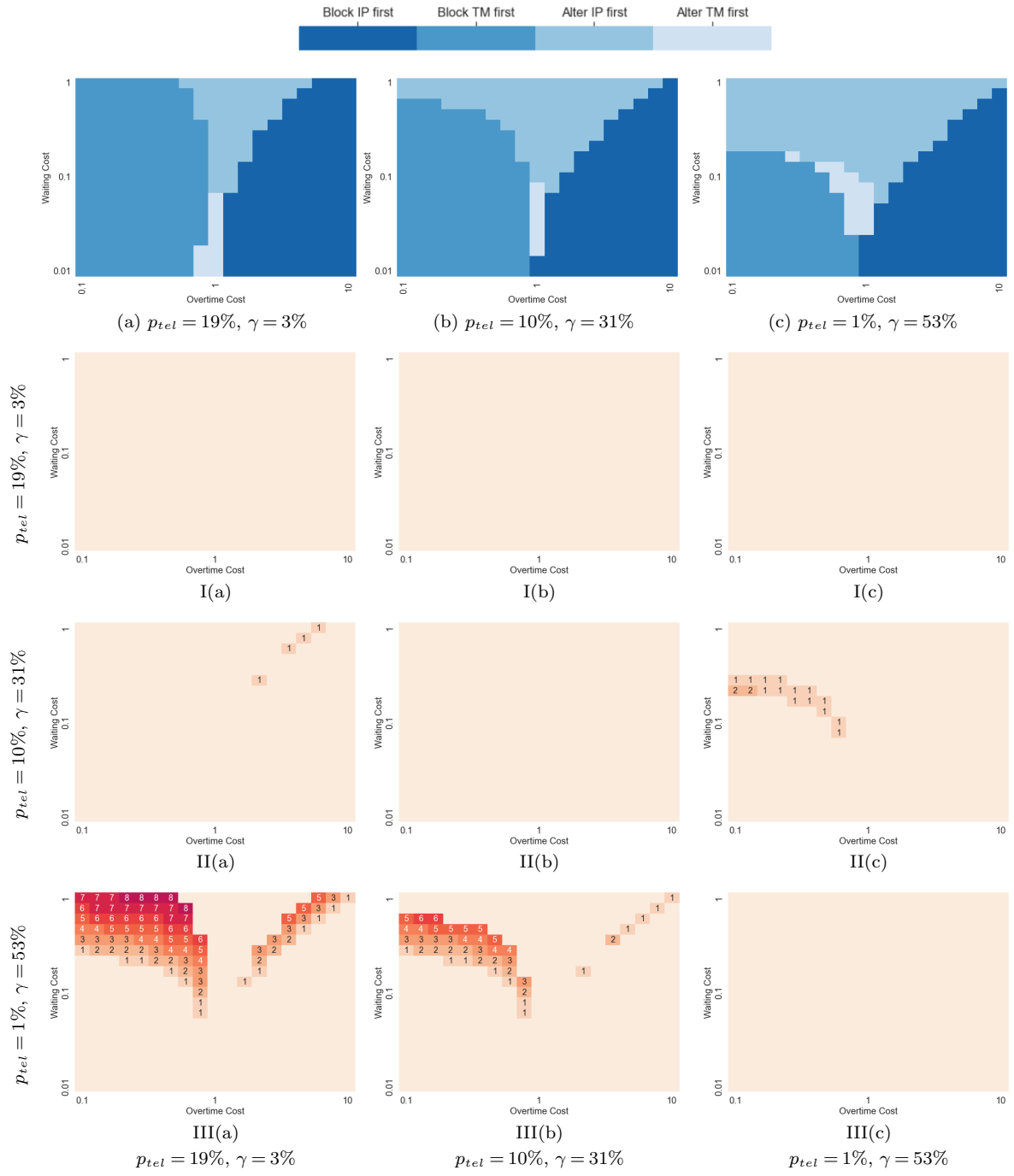


Figure 10 **First panel: best-performing heuristic sequencing rules. Second panel: the optimality gap in % when applying the policies derived based on the wrong system dynamics: each row of figures represents a specific system dynamic (under particular values of p_{tel} and γ), each column of figures represents the applying the policy depicted in the first panel to the system. $p_{in} = 20\%$. $\eta_{tel} = 20\%$. $T = 30$. $N = 42$.**

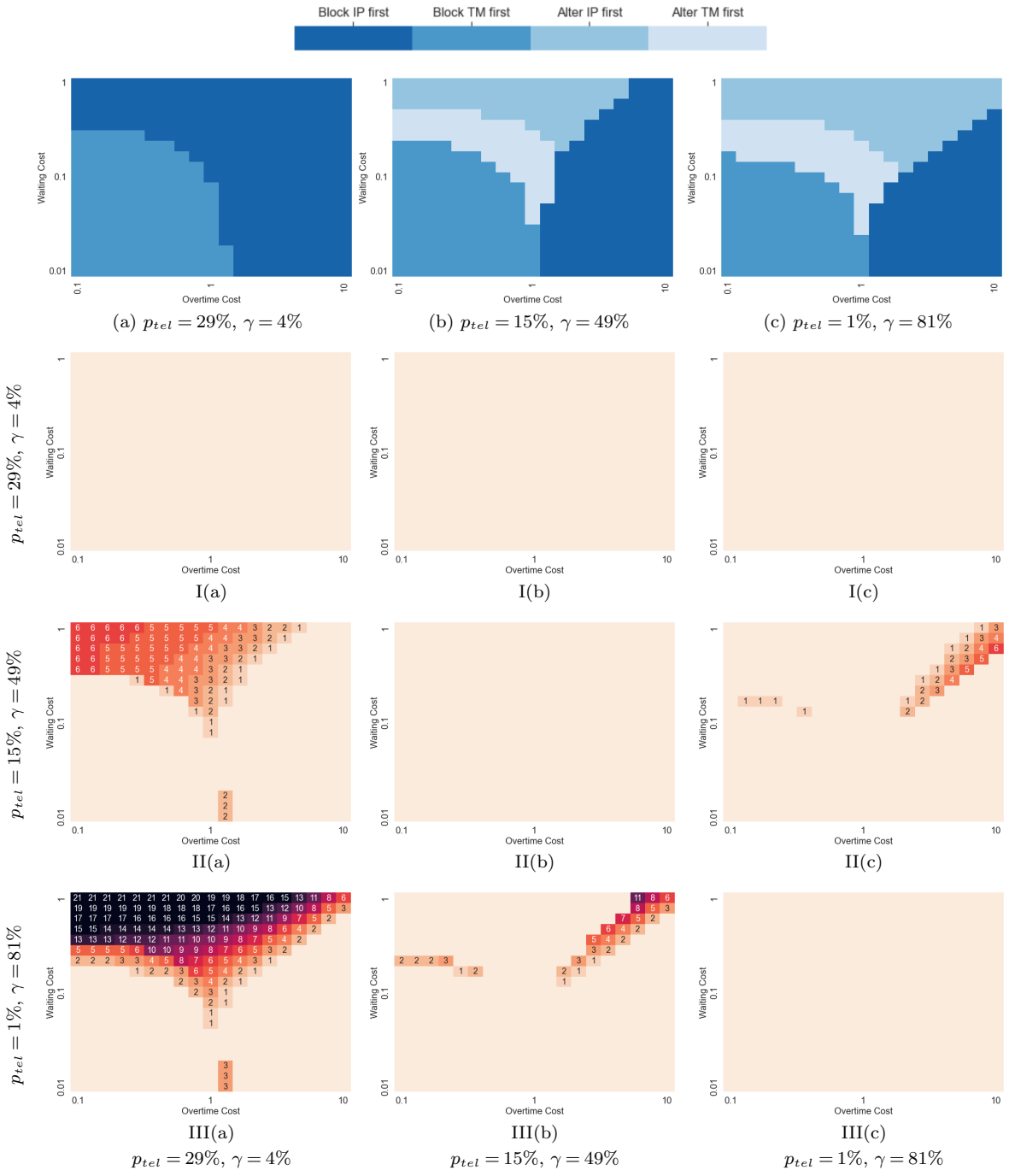


Figure 11 **First panel: best-performing heuristic sequencing rules. Second panel: the optimality gap in % when applying the policies derived based on the wrong system dynamics: each row of figures represents a specific system dynamic (under particular values of p_{tel} and γ), each column of figures represents the applying the policy depicted in the first panel to the system. $p_{in} = 30\%$. $\eta_{tel} = 30\%$. $T = 30$. $N = 36$.**

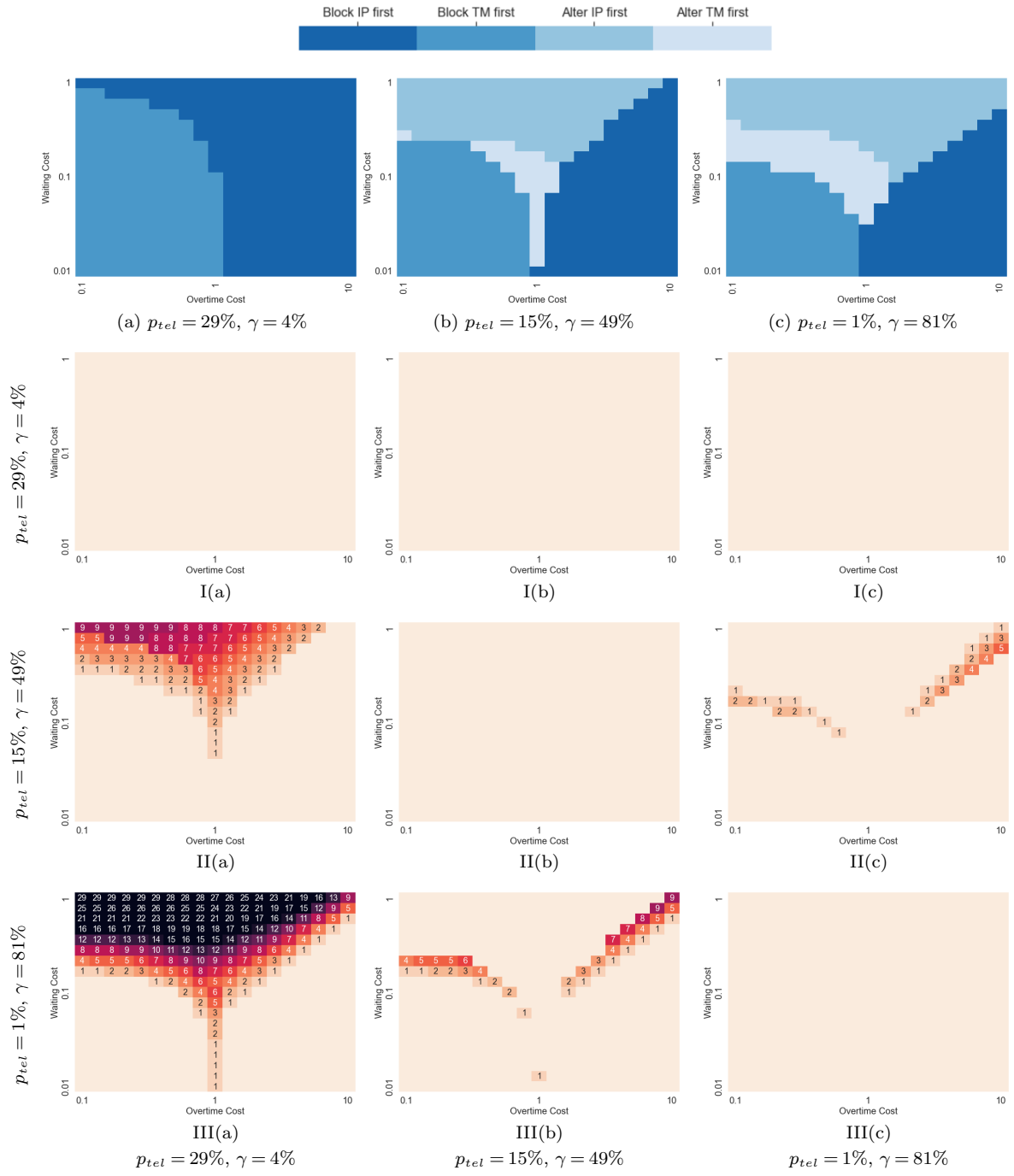


Figure 12 **First panel: best-performing heuristic sequencing rules. Second panel: the optimality gap in % when applying the policies derived based on the wrong system dynamics: each row of figures represents a specific system dynamic (under particular values of p_{tel} and γ), each column of figures represents the applying the policy depicted in the first panel to the system. $p_{in} = 30\%$. $\eta_{tel} = 30\%$. $T = 30$. $N = 42$.**

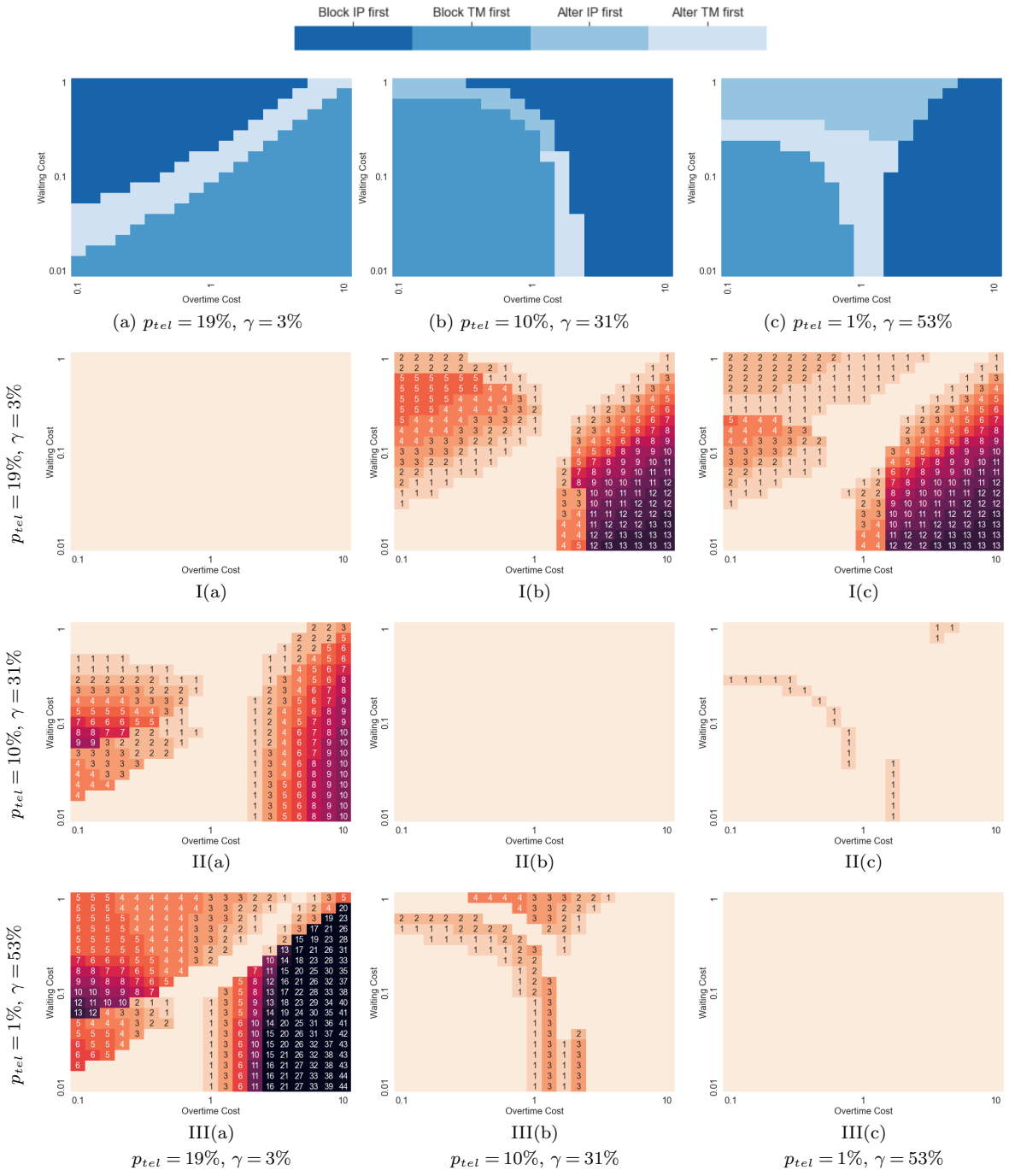


Figure 13 **First panel: best-performing heuristic sequencing rules. Second panel: the optimality gap in % when applying the policies derived based on the wrong system dynamics: each row of figures represents a specific system dynamic (under particular values of p_{tel} and γ), each column of figures represents the applying the policy depicted in the first panel to the system. $p_{in} = 30\%$. $\eta_{tel} = 20\%$. $T = 30$. $N = 42$.**

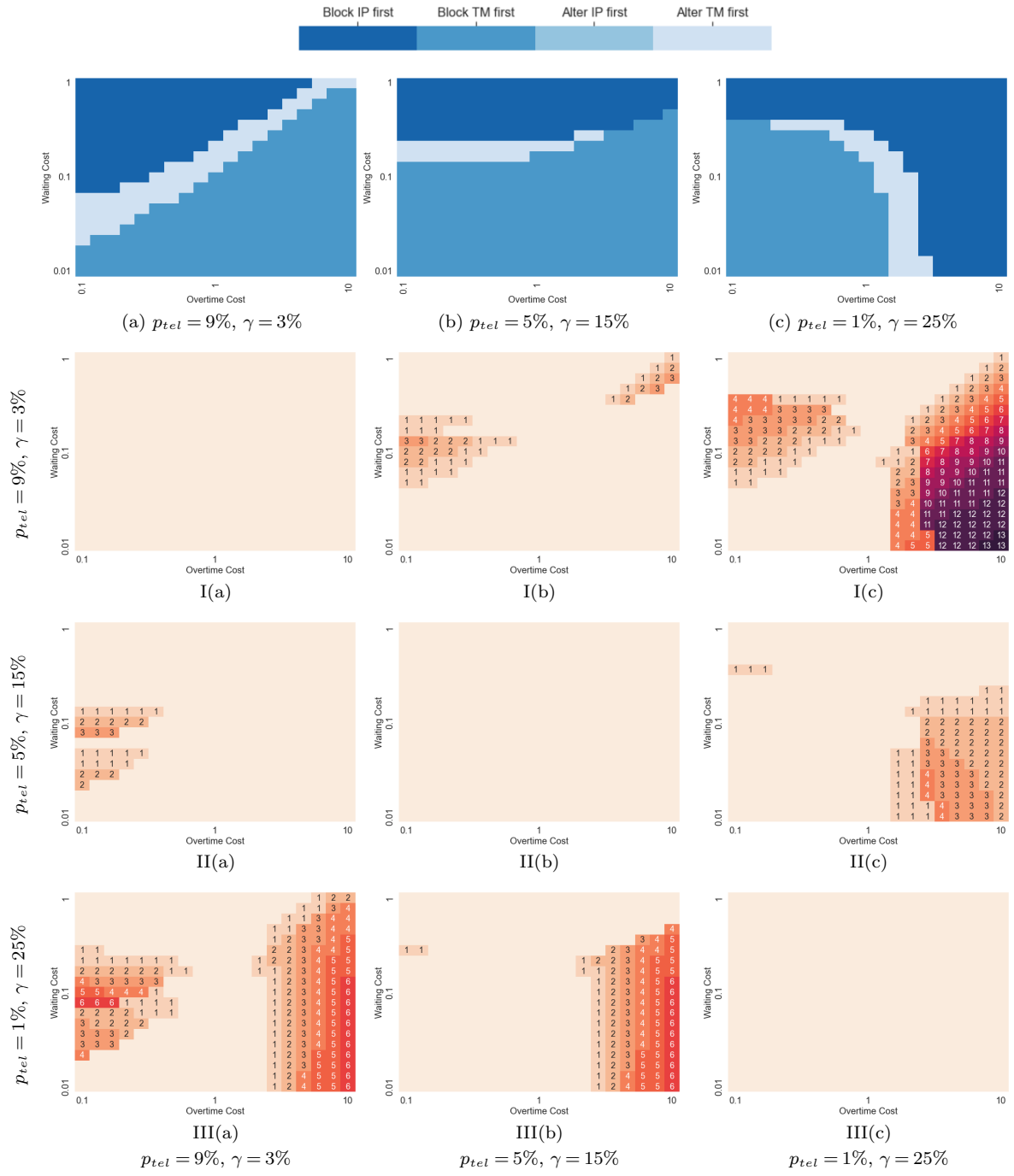


Figure 14 **First panel: best-performing heuristic sequencing rules. Second panel: the optimality gap in % when applying the policies derived based on the wrong system dynamics: each row of figures represents a specific system dynamic (under particular values of p_{tel} and γ), each column of figures represents the applying the policy depicted in the first panel to the system. $p_{in} = 20\%$. $\eta_{tel} = 10\%$. $T = 30$. $N = 36$.**

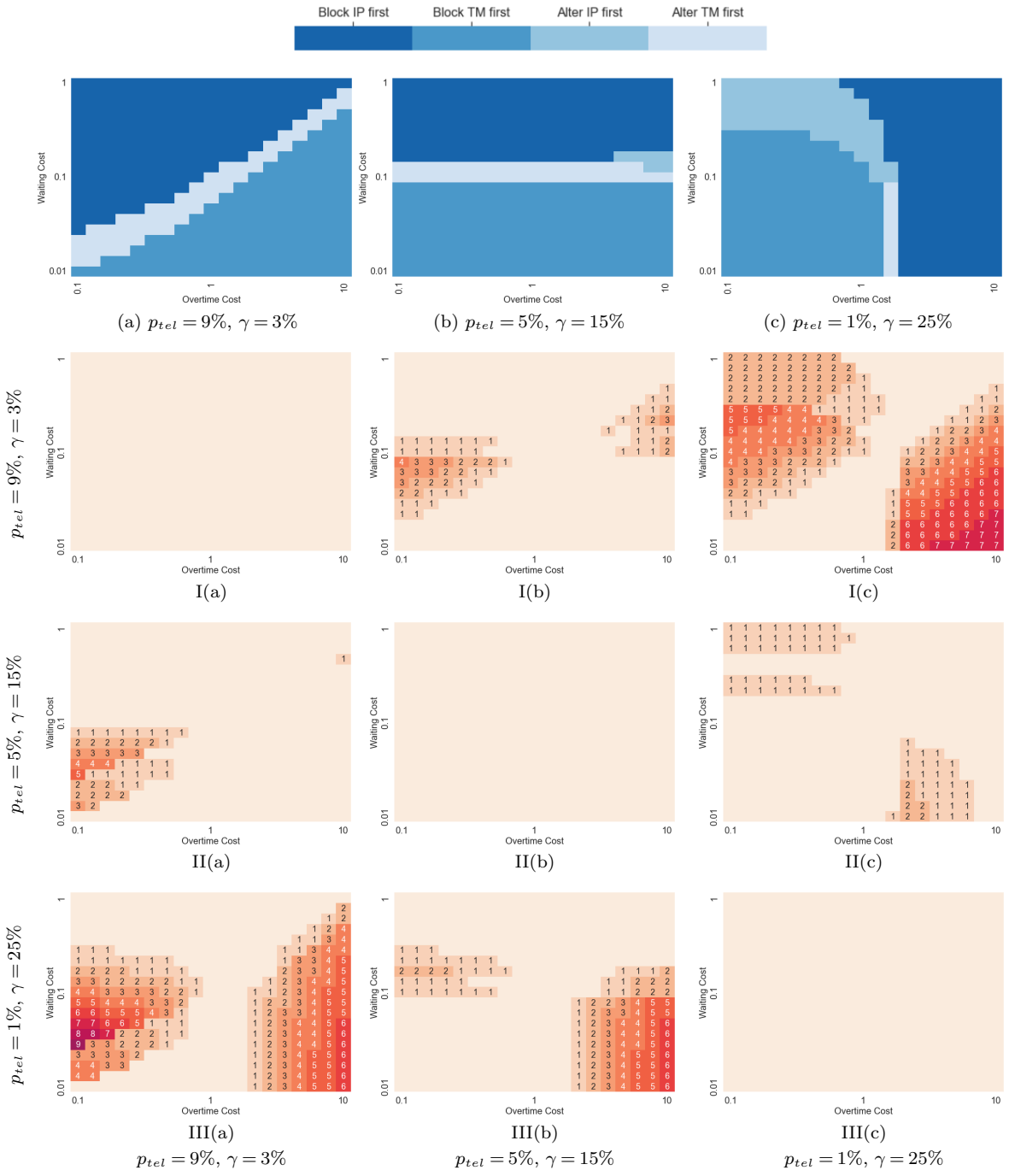
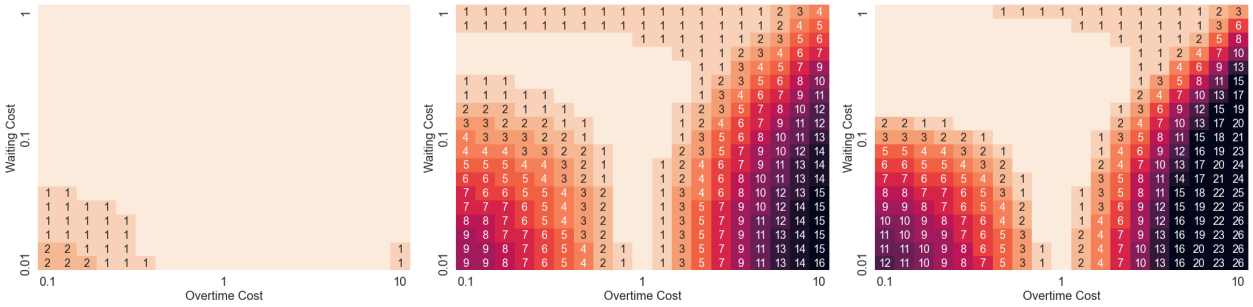
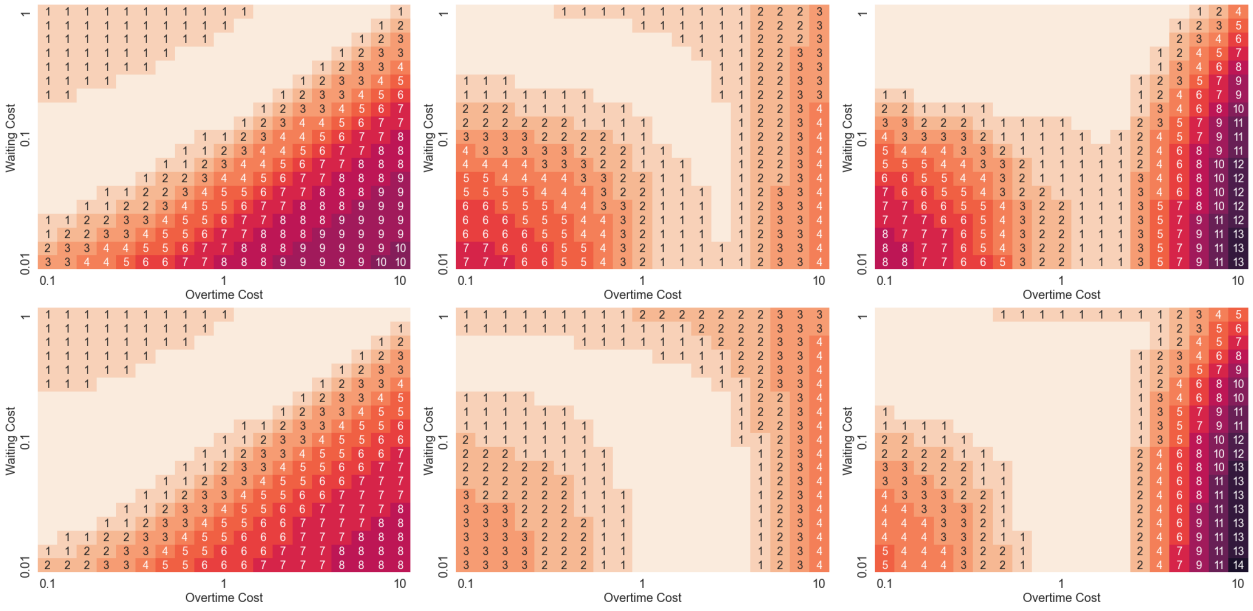


Figure 15 **First panel: best-performing heuristic sequencing rules. Second panel: the optimality gap in % when applying the policies derived based on the wrong system dynamics: each row of figures represents a specific system dynamic (under particular values of p_{tel} and γ), each column of figures represents the applying the policy depicted in the first panel to the system. $p_{in} = 20\%$. $\eta_{tel} = 10\%$. $T = 30$. $N = 42$.**



(a) $p_{tel} = 19\%$, $\gamma = 3\%$ (b) $p_{tel} = 10\%$, $\gamma = 31\%$ (c) $p_{tel} = 1\%$, $\gamma = 53\%$

Figure 16 **The optimality gap in % when applying the Alternating Telemedicine First policy.** $p_{in} = 20\%$. $\eta_{tel} = 20\%$. $T = 30$. $N = 36$.



(a) $p_{tel} = 19\%$, $\gamma = 3\%$ (b) $p_{tel} = 10\%$, $\gamma = 31\%$ (c) $p_{tel} = 1\%$, $\gamma = 53\%$

Figure 17 **The first (second) row shows the optimality gap in % when applying the Alternating In-Person First (Alternating Telemedicine First) policy.** $p_{in} = 30\%$. $\eta_{tel} = 20\%$. $T = 30$. $N = 36$.