The Impact of Information-Granularity and Prioritization on Patients' Care Modality Choices

Lin Zang
Simon Business School, University of Rochester, lzang@simon.rochester.edu

Yue Hu
Graduate School of Business, Stanford University, yuehu@stanford.edu

Ricky Roet-Green
Simon Business School, University of Rochester, ricky.roet-green@simon.rochester.edu

Shujing Sun
Jindal School of Management, University of Texas at Dallas, shujing.sun@utdallas.edu

The past few years have witnessed a significant expansion in telemedicine adoption by healthcare providers. On one hand, telemedicine has the potential to increase patients' access to medical appointments. On the other hand, due to the limitations of remote diagnostic and treatment methods, telemedicine may be insufficient for patients' treatment needs and may necessitate subsequent in-person follow-up visits. To better understand this tradeoff, we model the healthcare system as a queueing network providing two types of service: telemedicine and in-person consultations. We assume that an in-person visit guarantees successful treatment, whereas a telemedicine visit may fail to meet the patient's treatment needs with a probability that is contingent on individual patient characteristics. We formulate patients' strategic choices between these care modalities as a queueing game, and characterize the game-theoretic equilibrium and the socially optimal patients' choices. We further examine how improving patients' understanding of their telemedicine suitability through predictive analytics at the online triage stage affects system performance. We find that increasing information granularity maximizes the stability region of the system but may not always be optimal in reducing the average waiting time. This limitation, however, can be overcome by simultaneously deploying a priority rule that induces the social optimum under specific conditions. Finally, leveraging real-world data from a large academic hospital in the United States, we perform a comprehensive case study that encompasses both the development of a prediction model for in-person follow-up needs and the implementation of effective information provision and prioritization strategies.

Key words: Telemedicine, Online Triage, Strategic Queueing, Information Granularity, Waiting Times, Priority Rules

1. Introduction

In recent years, healthcare providers have extensively embraced telemedicine consultations, employing video and telephone technologies to deliver medical appointments to patients (Friedman AB and As 2022). Telemedicine presents the opportunity to diminish the risk of exposure to contagious diseases, enhance appointment accessibility, and alleviate the gap among patients from various socioeconomic groups (Kalwani et al. 2021, Sunar and Staats 2022, Osmanlliu et al. 2023, Qin et al. 2023a). However, its effectiveness is hindered by diagnostic and treatment limitations, potentially resulting in “duplicative care,” namely, the need for subsequent in-person follow-up visits when the telemedicine consultation fails to substitute for in-person treatment (Liu et al. 2021). Due to these advantages and drawbacks, both the healthcare literature and media discussions on the impact of integrating telemedicine frequently present a mix of positive and negative outcomes (Delana et al. 2023, SteelFisher et al. 2023).

The presence of duplicative care may lead to operational inefficiencies by increasing the overall workload of the system. However, addressing duplicative care following telemedicine encounters is challenging for two primary reasons. First, patients exhibit significant diversity in their illnesses, concurrent health conditions, and demographic characteristics. Even within primary care or the same (sub)specialty group, the appropriateness and efficacy of telemedicine treatment vary considerably among patients (Healthcare Dive 2022). Second, patients often lack the precise information and discernment necessary to determine whether a telemedicine consultation can adequately address their care needs. Consequently, when faced with a choice between care modalities—telemedicine or in-person visit—patients might unknowingly make decisions without sufficient information, opting for the less suitable visit type. Recognizing these challenges, health systems, as highlighted by Srinivasan et al. (2020), have acknowledged that “as the system expanded rapidly, providers also experienced frustration with scheduling patients for video visits rather than in-person visits when it was inappropriate, and they sought alterations to the triage and scheduling system.” Similarly, Kobeissi and Ruppert (2022) advocate for improved utilization of remote patient triage to ensure that telehealth effectively substitutes for in-person care among patients seeking virtual services.

Meanwhile, the growing accessibility of data and ongoing advancements in statistical learning techniques have presented a burgeoning opportunity to comprehend patient outcomes stemming from telemedicine visits. Considerable effort has been dedicated to constructing prediction models for various patient outcomes after telemedicine consultations, such as subsequent emergency department (ED) encounters and unplanned hospital readmissions (Shah et al. 2022, Hatef et al. 2022). Despite the increasing focus on predicting patient outcomes from telehealth, the question of how to effectively integrate predictive information into online triage tools and scheduling protocols remains relatively underexplored. The challenges faced in this context are twofold. First, most existing prediction models primarily concentrate on forecasting outcomes such as ED visits or hospitalizations. These models lack direct prediction capabilities regarding the likelihood of duplicative care, specifically, the need for an in-person follow-up visit (often with the same provider) after a telemedicine consultation. Second, while it is conceivable to develop a prediction model to forecast the probability of requiring an in-person follow-up after a telemedicine visit, the translation of this additional predictive information into improved system performance—such as reducing average waiting times or increasing patient throughput—remains unclear. Moreover, since patients are the primary decision-makers in most ambulatory clinic settings, their choices have a cascading effect, potentially influencing the decisions of other patients. Consequently, it is not immediately evident whether providing more predictive information to patients through an online triage tool will enhance overall system performance.

In this paper, we develop a comprehensive framework to integrate predictive analytics into online triage and scheduling protocols. Our theoretical model centers on a queueing-game model that encompasses patients' choices between two distinct care modalities: telemedicine consultations and in-person visits. We focus on application settings where the objective of deploying telemedicine services is to substitute for in-person visits, although this substitution may not be successful for some patients. To capture the heterogeneity in patients' suitability for telemedicine treatment, we assume that the likelihood of a patient requesting an in-person follow-up after a telemedicine consultation is a random variable, namely, a function of random patient type.

To incorporate predictive analytics into our theoretical framework, we construct a prediction model leveraging data obtained from a large academic hospital in Maryland, United States. In January 2021, the hospital started offering telemedicine alternatives for a range of preprocedural assessment, where remote planning are used to substitute for in-person evaluation. Similar practices have been documented by Mihalj et al. (2020), Crawford et al. (2021), and Eyrich et al. (2022). Utilizing encounter-level outpatient data, we develop a logistic regression model to forecast the probability that a patient will need an in-person follow-up visit subsequent to remote consultation. Furthermore, to embed the prediction model into the theoretical framework, we assume that for each patient, the prediction model is able to predict the realized (as opposed to random) probability of needing an in-person follow-up after telemedicine treatment.

Our study aims to evaluate the potential benefits of offering predictive information to patients through an online triage tool, utilizing both the prediction model and the queueing-game model. We specifically analyze the effects of two information granularity regimes—crude and refined—on system performance. Under the crude information granularity regime, patients lack precise information about their specific (realized) types and, consequently, their probability of requiring in-person follow-up after telemedicine treatment. Here, a patient bases their choice between care modalities on the distribution of in-person follow-up needs after a telemedicine consultation. In contrast, in the refined information granularity regime, the prediction model is integrated into the online triage tool. Upon entering the system, the patient is informed of their realized type and the probability that they will need in-person follow-up care as suggested by the prediction model. In both information granularity regimes, patients, conditioned on the available information, make decisions to minimize their expected total waiting time in the system. This total waiting time encompasses both the waiting time for the initial care and the potential subsequent waiting time in the event of needing an in-person follow-up visit. Our main contributions can be summarized as follows.

Impact of Information Granularity on Patients' Behavior. We use the queueing-game and prediction models to characterize mixed-strategy Nash equilibria in crude and refined information regimes. By comparing system performance across these two information setups, we assess the benefits of offering more detailed information to patients via an online triage tool. First, in terms of system stability, our findings consistently demonstrate the benefits of providing enhanced information to patients. Specifically, the refined information granularity regime achieves the maximal parameter space within which both the in-person and telemedicine queues remain stable under certain mixed-strategy Nash equilibria. Second, in terms of average waiting time, our analysis uncovers that in the refined information granularity regime, the average waiting time may be lower than, equivalent to, or, surprisingly, higher than that in the crude information regime. While the scenario where increased information leads to longer waiting times is confined to a relatively small parameter region, this result underscores the need for providers to exercise caution when using remote triage to influence patients' behavior.

Mechanism to Achieve the System's First Best Performance. To gain deeper insights into the impact of the online triage tool in the refined information granularity regime, we conduct detailed analysis of the optimal patient routing in a context where decision making is centralized (as opposed to strategic decision making by the patients), referred to as the system's "first best solution." This first best framework establishes a benchmark for the lowest possible average waiting time under patients' strategic decisions. We find that in certain parameter regions, the average waiting time under patients' equilibrium strategy in the refined information regime is identical to that under the system's first best solution. In this regime, it is notable that providing more information to guide patients' behavior already achieves the optimal centralized performance. Additionally, in parameter regions where the system's first best solution surpasses patients' choices, we introduce a priority rule that incentivizes patient preferences through prioritization. This policy effectively steers the system's performance towards the first best solution. Importantly, in the seemingly "problematic" instances where the refined information granularity initially yields a longer average waiting time than the crude information regime, we demonstrate that employing the priority rule can transform this information disadvantage into an advantage. That is, through the joint provision of information and appropriate prioritization, the system indeed attains the first best performance.

Practical Insights and Implementation Framework. To facilitate real-world implementation, we complement the theoretical model by leveraging real-world data from a large academic hospital in Maryland. Our approach involves two key steps: 1) developing a prediction model to forecast the need for in-person follow-up after telemedicine consultations, and 2) conducting a comprehensive case study to demonstrate model calibration and performance evaluation in a real-world setting. To the best of our knowledge, we make a pioneering effort in constructing a prediction model to forecast duplicative care based on real-world data. Based on the case study, we estimate that providing patients with information through online triage can lead to a remarkable reduction in average waiting time from 14.48 days to 8.56 days, reflecting a substantial 41.21% decrease. The average waiting time can be further reduced to 8.13 days (a further 4.99% decrease) by offering proper prioritization among patients. Importantly, our generic framework can be easily adopted by other health providers to assess the impact of operational levers, particularly in information design and scheduling policy, on overall system performance.

1.1. Related Literature

Telemedicine Adoption and Patients' Behavior. Our work is related to the healthcare operations management literature that studies how to better manage telemedicine services.

Many papers utilize empirical methods to study patients' behavior and assess the impact of telemedicine. For example, Qin et al. (2023b) estimate the causal effect of physician availability on service incompletion rates. Bavafa et al. (2018) and Lekwijit et al. (2023) analyze the impact of telemedicine services on the utilization rates of in-person services. Other works evaluate the effect of telemedicine on other outcomes such as patient adherence and access expansion (Staats et al. 2017, Li et al. 2021, Sun and Wang 2021, Delana et al. 2023).

In addition to empirical investigations, several studies develop analytical models to optimize operational decisions, such as in managing workload (Saghafian et al. 2018) and determining visit intervals (Bavafa et al. 2021). Notably, a subset of research employs game-theoretic and queueing approaches to 1) model the strategic behavior of patients in selecting care modalities, and 2) analyze and optimize the resulting game-theoretic equilibrium within the system. For instance, Çakıcı and Mills (2022) utilize a three-stage game-theoretic model to examine the effects of reimbursement policies for both telemedicine and in-person visits. Meanwhile, Rajan et al. (2019) apply a queueing-game model to investigate how telemedicine technology influences patient utility and pricing decisions. Ding et al. (2022) examine the influence of binary recommendation (for ED vs.

general practitioner) on patients' behavior with partially endogenous queueing dynamics. Additionally, Liu et al. (2023) use a queueing-game model to study the optimal allocation of service capacity between different care modalities. Similarly, our paper adopts an analytical framework that combines queueing and game theory. Despite these parallels, we make a novel and important contribution to the existing literature by focusing on the joint impact of information design and priority rules on system performance. Our unified framework encompasses the development of 1) a prediction model for forecasting in-person follow-up needs after tele-consultation, 2) an information provision strategy for conveying predictive information to patients, and 3) a priority rule that helps orient patients' equilibrium towards achieving the system's first best performance.

Information Design for Service Systems. Extensive research has been conducted to examine the value of information and develop effective strategies for information provision in service systems. A substantial body of literature considers whether and how to communicate delay announcements to customers; see, e.g., Akşin et al. (2017), Yu et al. (2017, 2022). We refer interested readers to Ibrahim (2018) for a comprehensive review. In contrast, our study focuses on the practice of providing the "predicted suitability/efficacy" of telemedicine consultations for patients, introducing a novel modeling context distinct from those employed in the investigation of delay announcements. Despite this distinction, our findings reveal interesting parallels with the delay announcement literature in terms of the managerial insights yielded. As pointed out by Ibrahim (2018), heterogeneity can be exploited through delay announcements. In our work, we leverage the heterogeneity in patients' suitability for telemedicine through information provision and priority rules, resulting in variations in expected waiting times and fostering incentives for optimal patient choices. Furthermore, aligning with existing research on delay announcement, our study underscores the insight that more information is not always better. Specifically, we characterize the parameters regimes where providing more information to patients paradoxically leads to an increase in average waiting times. Consequently, healthcare providers must exercise discretion in designing the information disclosed to patients during online triage processes.

Our research is also related to the class of literature addressing the integration of predictive information into operational decisions. For example, Argon and Ziya (2009), Sun et al. (2022), Singh et al. (2024) investigate the utilization and design of prediction models for customer classification in multi-class queueing networks. Hu et al. (2022b) explore how to schedule proactive care based on predicted patient deterioration. Moreover, Hu et al. (2023) develop a two-stage staffing framework driven by demand predictions for ED nurse staffing. While sharing common goals with these studies in bridging predictive analytics and prescriptive operational decisions, our work distinguishes itself by focusing on strategic agent behavior and interactions.

Strategic Behavior in Queueing Systems. Our work aligns with the queueing literature that accounts for customers' strategic behavior and rational decision making; see, for example, Naor (1969), Edelson and Hilderbrand (1975), Hassin (1986), Hassin and Haviv (2003), Hassin (2016), Hassin and Roet-Green (2017, 2021). The majority of existing literature focuses on the use of pricing as a control lever. A number of papers further incorporate priority into the queueing context, considering how the service provider should determine entry prices for priority queues, run priority auctions, and offer different priorities contingent on customers' disclosure of information (Mendelson and Whang 1990, Afeche and Mendelson 2004, Afeche et al. 2019, Hu et al. 2022a). A thorough review of research on priority queues with self-interested customers can be found in Cui et al. (2023). In contrast with the aforementioned settings, our work considers the joint levers of information design and priority rules. We examine the impact of incorporating predictive analytics into the queueing-game framework. These novel mechanisms are particularly pertinent to healthcare settings where providers often lack direct control over setting payments for patients. We subsequently evaluate the performance disparity between the game-theoretic equilibrium under our proposed policies and the steady state under optimal centralized control. This unique perspective sheds light on the interplay between predictive analytics, strategic behavior, and queueing dynamics.

Additionally, the tradeoff in service quality between two distinct types of servers, the gatekeeper and the specialist, is similarly captured by the well-known gatekeeper's model, as examined by Shumsky and Pinker (2003), Freeman et al. (2017), Hathaway et al. (2023). Specifically, the gatekeeper model literature considers a two-tier service system where customers initially enter the system through the gatekeeper (first-tier server). The gatekeeper then makes strategic decisions in transferring customers to the specialist (second-tier server). In contrast, our paper considers a two-channel service system without hierarchical "tiers." Instead of joining the gatekeeper queue first, patients in our model make strategic choices between care modalities upon arrival. Consequently, the pivotal agent with strategic decision-making in our model is each patient, as opposed to the gatekeeper in the gatekeeper framework. A similar two-channel system has been presented by Roet-Green and Shetty (2022), who model regular and expedited airport security check lines. However, their focus is on assessing the impact of fees and resource allocation on customer decision-making, with an emphasis on fairness. In contrast to our paper, they do not consider the likelihood of unsuccessful substitution between service channels, which is the main motivation of our model.

1.2. Organization of The Paper

The rest of the paper is organized as follows. In Section 2, we introduce the model and the two information granularity regimes. We characterize the game-theoretic equilibria of the model and compare system performance across different information granularity regimes in Section 3 and Section 4. In Section 5, we derive the system's first best solution and analyze the disparity between the equilibrium under patients' strategies and the steady state under the optimal centralized solution. Then in Section 6, we explore the use of priority rules in conjunction with information provision as a mechanism to orient the patient's equilibrium towards the first best performance. In Section 7, we conduct a comprehensive case study that includes the construction of a prediction model, calibration of parameters, and application of theoretical results in a real-world setting. Our conclusion and future directions are presented in Section 8.

2. The Model

We consider a queueing system (depicted in Figure 1) with continuously distributed patient types and two distinct provider care modalities—telemedicine and in-person visits—in which providers differ in their treatment competency. Patients arrive to the system according to a Poisson process with rate $\lambda \in \mathbb{R}_+$ and with their types uniformly distributed over the real interval $[0,1]$. To capture heterogeneous patients' suitability for telemedicine visits, we let $f:[0,1]\to[0,1]$ be a mapping between the patient type and the likelihood of not receiving sufficient treatment through telemedicine. Specifically, for a patient of type $t \in [0,1]$, a telemedicine visit fails to fulfill their treatment requirement with probability $f(t)$, upon which the patient needs an in-person follow-up visit and joins the queue for the in-person provider. We assume that f is strictly increasing. That is, a patient with a larger type value has a higher chance of encountering an unsatisfactory treatment episode through telemedicine. In addition, we assume that $f(0)=0$ and $f(1)=1$. Namely, with probability 1, the "least severe" patient (of type 0) can be successfully treated by telemedicine, and the "most severe" patient (of type 1) needs an in-person follow-up after telemedicine visits. Moreover, we assume a fully Markovian system, where the service rates of telemedicine and in-person visits are $\mu_1, \mu_2 \in \mathbb{R}_+$, respectively. Within each queue (for telemedicine or in-person visits), patients are treated on a first-come-first-served basis (with the exception of Section 6, where priority rules are considered).

By model construction, we assume that the probability of incomplete telemedicine treatment is the random variable $X = f(U)$ for $U \sim \text{Uniform}[0,1]$. This model formulation captures a wide range of probability distributions for X due to the flexibility in the functional form of f. In particular, let F_X denote the cumulative distribution function of X. Based on the inverse transform sampling method, it holds that $X \overset{d}{=} f(U)$ for $f := F_X^{-1}$. Hence, our assumptions on f are equivalent to assuming that F_X^{-1} is continuous and strictly increasing.

Patients are self-interested and rational utility maximizers who, upon arrival, choose to join either the queue for telemedicine visits or the queue for in-person visits based on the expected steady-state utilities. To model patient utility, we let $R \in \mathbb{R}_+$ denote the value of service and $c \in \mathbb{R}_+$ denote the cost of waiting per unit time in the queue for each patient. Then a patient's net utility from receiving service is given by

$$R - c \cdot \mathbb{E}[W],$$

where W is the steady-state total waiting time that includes both the initial waiting time for the telemedicine or in-person visit, and, if applicable, the subsequent waiting time for the in-person follow-up visit.

We conclude this subsection by noting that we do not incorporate patients' balking behavior into the model. That is, we assume that R is sufficiently large so that patients receive non-negative utility through either an in-person or a telemedicine visit, aside from the probability of needing in-person follow-up after insufficient telemedicine treatment. This assumption adds complexity to the stability conditions of the model, as the stability of the queues relies on patients' strategic behavior induced by the information design and priority rules; see our formal definition in Section 2.3. Our subsequent analysis aims to shed light on how information provision and priority rules impact the parameter space in which the system maintains stability.

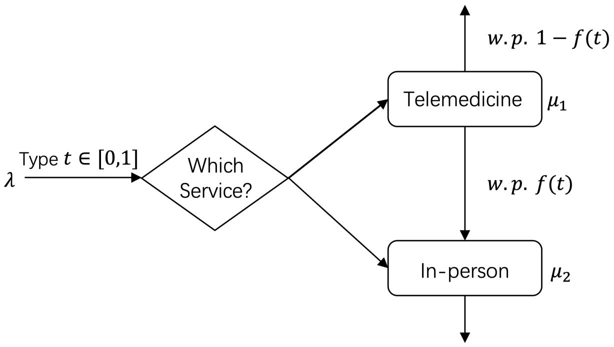


Figure 1 Patients' Decision and Flow

2.1. Information Granularity Regimes

To investigate the impact of information on patients' choices and system performance, we consider two information granularity regimes, which we refer to as "crude" and "refined," respectively.

In the crude information regime, a patient has knowledge of the probability distribution associated with the likelihood of needing in-person follow-ups after a telemedicine visit, specifically, the distribution of X. However, patients are not privy to their individual realizations of X. Recall that $X = f(U)$, where $U \sim \text{Uniform}[0,1]$. This assumption equivalently implies that patients are aware of the functional form of f but are unaware of their specific realized patient types.

In contrast to the crude information regime, the refined information granularity regime assumes that patients have knowledge not only of the probability distribution of X but also its specific realization linked to their individual care episode. In other words, patients are aware of not only the functional form of f but also of their realized patient types.

2.2. Patients' Strategies

We allow the set of admissible patient strategies to be type-dependent mixed strategies. In particular, a patient of type $t \in [0,1]$ chooses telemedicine consultation with probability $g(t)$ and selects in-person visit with probability $1 - g(t)$, for some mapping $g : [0,1] \to [0,1]$. Given patients' strategies g, the steady-state total waiting time for patient of type t is given by

$$W(g,t) := \mathbf{1}_{Tele}(g(t))\big(W_1(g) + \mathbf{1}_{Follow-up}(t) W_2(g)\big) + \big(1 - \mathbf{1}_{Tele}(g(t))\big) W_2(g), \tag{1}$$

and the average (across all patients) steady-state total waiting time in the system is

$$\mathbb{E}\left[W(g,U)\right]. \tag{2}$$

In Equation (1) above, the indicator function $\mathbf{1}_{Tele}$ denotes whether the patient chooses telemedicine service. The indicator function $\mathbf{1}_{Follow-up}$ denotes whether an in-person follow-up visit is needed after the telemedicine consultation. In addition, W_1 and W_2 are steady-state waiting times at the telemedicine queue and the in-person queue, respectively. To better understand how the objective is calculated, note that the first term in Equation (1) corresponds to the total steady-state waiting time in the system after choosing the telemedicine service, which consists of both the waiting at the telemedicine queue and the subsequent waiting at the in-person queue if an in-person follow-up is needed. In analogue, the second term in Equation (1) corresponds to the total steady-state waiting time in the system after choosing the in-person service, which is simply the steady-state waiting time at the in-person queue. We next define each patient's self-interested optimization problem to determine his/her joining strategy in the two systems respectively.

In the crude information regime, a patient of type U sets the probability of using the telemedicine service, taking into account the randomness in his/her type (and thus in the probability of needing an in-person follow-up after telemedicine). In particular, given the other patients' strategies $g: [0,1] \to [0,1]$, the self-interested optimization problem for patient of type U is given by

$$\min_p \mathbb{E}_U\left[\mathbf{1}_{Tele}(p)\big(W_1(g_{-U},p) + \mathbf{1}_{Follow-up}(p)W_2(g_{-U},p)\big) + \big(1 - \mathbf{1}_{Tele}(p)\big)W_2(g_{-U},p)\right], \tag{3}$$

where $W_1(g_{-U},p)$ denotes the steady-state waiting time in the telemedicine queue given that a patient of type U chooses telemedicine with probability p and other patients (of type t for $t \neq U$) select telemedicine based on strategies mapped by the correspondence g. Note that the expectation in Equation (3) is taken with respect to the following three sources of uncertainty: the type of the focal patient, the patient's choice of telemedicine or in-person visit, and queueing dynamics. We add the subscript in $\mathbb{E}_U$ to highlight the uncertainty in the patient type U, which is in contrast to the patient's self-interested optimization in the refined information regime that we introduce next.

In the refined information regime, patients know the realization of their types and thus the corresponding probabilities of needing in-person follow-up visits after tele-consultation. In particular, given the other patients' strategies $g: [0,1] \to [0,1]$, the self-interested optimization problem for a patient of type $t \in [0,1]$ is given by

$$\min_p \mathbb{E}\left[\mathbf{1}_{Tele}(p)\big(W_1(g_{-U},p) + \mathbf{1}_{Follow-up}(U)W_2(g_{-U},p)\big) + \big(1 - \mathbf{1}_{Tele}(p)\big)W_2(g_{-U},p) \,\middle|\, U = t\right], \tag{4}$$

where the expectation in Equation (4) is taken with respect to the selection decision of telemedicine and queueing dynamics, conditioned on the type of the focal patient.

In either information granularity regime, we assume that patients with identical perceived type information employ the same strategy. In Section 3 below, we characterize the unique mixed strategy Nash equilibrium for the patients' self-interested problems in Equations (3) and (4). We then compare the equilibrium system performance in the crude and refined information regimes in terms of 1) the expected steady-state total waiting time (characterized in Equation (2)) under the equilibrium patient strategies, and 2) the system stability region, which we introduce in the next subsection.

2.3. System Stability Region

In our model, the conditions to ensure system stability are highly non-trivial because 1) there exist possible needs for in-person follow-up visits after telemedicine treatment, 2) patients do not balk or abandon, and 3) the stability of the queues depends on the patient strategy g. For a given patient strategy g, we say that the queueing system is stable if and only if it admits a well-defined steady state for the joint queue length process.

Before characterizing how the system stability condition is influenced by g, we first discuss two necessary conditions for system stability in Assumption 1, which is assumed throughout the paper.

Assumption 1 *The following conditions are necessary for system stability: 1) $\mu_1 + \mu_2 > \lambda$, and 2) $\mu_2 > \lambda\mathbb{E}[X]$.*

To interpret Assumption 1, note that the first condition ensures system stability when telemedicine services can fully substitute for in-person visits, namely, there need not be any in-person follow-up visits after telemedicine consultations. For the second condition, note that for a given patient strategy g, the arrival rate at the in-person visit is $\lambda(\int_0^1 (g(x)f(x) + 1 - g(x))dx)$, where the integrand is decreasing in $g(x)$, $x \in [0,1]$. In addition, since $g(x) \leq 1$ for all $x \in [0,1]$, it is straightforward to see that the arrival rate for the in-person queue is lower bounded by $\lambda(\int_0^1 f(x)dx)$. Therefore, a necessary stability condition is to have $\mu_2 > \lambda(\int_0^1 f(x)dx) = \lambda\mathbb{E}[f(U)] = \lambda\mathbb{E}[X]$.

We next characterize the system stability condition for any arbitrary patient strategy g.

Lemma 1 (System stability condition) *Given patient strategy $g : [0,1] \to [0,1]$, the system possesses a well-defined steady state if and only if:*

$$\mu_1 - \lambda \int_0^1 g(x)dx > 0$$

and

$$\mu_2 - \lambda \left(\int_0^1 (g(x)f(x) + 1 - g(x))dx \right) > 0.$$

Finally, for a specified information granularity regime, we say that the queueing system is stable in Nash equilibrium if there exists a mixed strategy g such that 1) the system is in mixed strategy Nash equilibrium under g, and 2) the system is stable with respect to g. We define $\mathcal{S}^{crude} := \{(\lambda, \mu_1, \mu_2, f) \in \mathbb{R}_+^3 \times C[0,1]\}$ and $\mathcal{S}^{refined} := \{(\lambda, \mu_1, \mu_2, f) \in \mathbb{R}_+^3 \times C[0,1]\}$ such that for each quadruple of model primitives $(\lambda, \mu_1, \mu_2, f) \in \mathcal{S}^j$ where $j \in \{crude, refined\}$, the system is stable under Nash equilibrium. As mentioned, to facilitate the comparison of information granularity regimes, one of the metrics we consider is the system stability region, which essentially involves comparing $\mathcal{S}^{crude}$ and $\mathcal{S}^{refined}$. Explicit characterization of the stability regions $\mathcal{S}^{crude}$ and $\mathcal{S}^{refined}$ is provided in Corollaries 1 and 2 in Section 3 after we analyze patients' strategies and system equilibria. Here, we refer to $\mathcal{S}^{crude}$ and $\mathcal{S}^{refined}$ as "system stability regions," diverging slightly from conventional terminology. While traditional queueing literature typically emphasizes the stability of queues under specific strategies, our criteria extend beyond queue stability to ensure the attainment of game-theoretic equilibrium. Therefore, the stability region encompasses both the conventional notion of stability in queueing theory and the requirement for equilibrium in game theory. It is noteworthy that these conditions may overlap, as is the case in our model; see details in Section 3 below.

3. Patient Equilibrium

In this section, we characterize patients' equilibrium strategies and system performance in the crude and refined information granularity regimes, respectively.

3.1. The Crude Patient Equilibrium

In the crude information regime, patients know the functional form of f but do not know the realization of their type U. The corresponding self-interested optimization problem for a patient of type U is formulated in Equation (3). Recall that by assumption, patients with identical perceived type information employ the same strategy. In the crude information regime, all patients have the same information, and thus, they apply the same mixed strategy of selecting telemedicine. Hence, it is without loss of generality to restrict to patient strategies of the form $g:[0,1]\to[0,1]$ such that $g(U)=p$, for some $p\in[0,1]$.

By Lemma 1, there exists a strategy in the aforementioned family of patient strategies parameterized by a single probability under which the system is stable if and only if

$$\mu_1-\lambda p>0 \quad \text{and} \quad \mu_2-\lambda(1-p+p\mathbb{E}[X])>0 \quad \text{for some } p\in[0,1],$$

which, by basic mathematical rearrangements, can be equivalently expressed as

$$\mu_2>\lambda-\mu_1+\mu_1\mathbb{E}[X]. \tag{5}$$

Assuming that Equation (5) holds, we say that the system is in mixed strategy Nash equilibrium under patient strategy $g^*:[0,1]\to[0,1]$ if and only if there exists $p^*\in[0,1]$ such that

$$g^*(U)=p^* \tag{6}$$

and

$$p^*=\arg\min_{p}\,\mathbb{E}_U\left[\mathbf{1}_{Tele}(p)(W_1(g^*_{-U},p)+\mathbf{1}_{Follow-up}(U)W_2(g^*_{-U},p))+(1-\mathbf{1}_{Tele}(p))W_2(g^*_{-U},p)\right]. \tag{7}$$

The next proposition establishes the existence of a unique mixed strategy Nash equilibrium when patients' information granularity is crude.

Proposition 1 *Suppose that Equation* (5) *holds. In the crude information granularity regime, there exists a unique mixed strategy Nash equilibrium that satisfies conditions* (6) *and* (7) *with* p^* *characterized as follows:*

1. *When* $\lambda + (1-\mathbb{E}[X])\mu_1 \leq \mu_2$*, all patients select the in-person visit, i.e.,* $p^* = 0$*;*
2. *When* $\mu_1 > \lambda$ *and* $\mu_2 \leq (1-\mathbb{E}[X])\mu_1 - (1-2\mathbb{E}[X])\lambda$*, all patients choose the telemedicine service, i.e.,* $p^* = 1$*;*
3. *Otherwise, each patient chooses the telemedicine service with probability*

$$p^* = \frac{(1-\mathbb{E}[X])\mu_1 + \lambda - \mu_2}{2\lambda(1-\mathbb{E}[X])}.$$

Based on the expression of p^* in Proposition 1, it can be derived that the crude patient equilibrium p^* increases with respect to μ_1 and decreases with respect to μ_2 (Lemma B.1 in Appendix B.4). Moreover, recall that Equation (5) presents the condition under which the queues are stable under some probabilistic type of patient strategy. Proposition 1 further establishes that the system attains a unique game-theoretic equilibrium under this strategy. As formalized in the following corollary, it immediately follows that $\mathcal{S}^{crude}$ coincides with the parameter space characterized by Equation (5).

Corollary 1 *It holds that* $\mathcal{S}^{crude} = \{(\lambda, \mu_1, \mu_2, f) : \mu_2 > \lambda - \mu_1 + \mu_1\mathbb{E}[X]\}$.

3.2. The Refined Patient Equilibrium

In the refined information regime, patients know not only the functional form of f but also the realization of their type U. The corresponding self-interested optimization problem for a patient of type $U = t$ is presented in Equation (4). Since the probability of needing in-person follow-up visits after telemedicine treatment, captured by the function f, is strictly increasing, it is intuitive to expect patients' equilibrium mixed strategies g to be decreasing, namely, $g(t_1) \leq g(t_2)$ for $t_1 \leq t_2$. The next lemma formalizes this intuition and further establishes that the equilibrium mixed strategies must be of a threshold type.

Lemma 2 *In the refined information granularity regime, if* $g : [0,1] \to [0,1]$ *is an equilibrium mixed strategy, then* g *is threshold-type. That is, there exists some threshold* $t^* \in [0,1]$ *such that* $g(t) = 1$ *for* $t \leq t^*$*, and* $g(t) = 0$ *for* $t > t^*$*.*

Based on Lemma 2, it is without loss of generality to restrict attention to the family of threshold-type patient's strategies. Furthermore, by Lemma 1, there exists a threshold-type strategy under which the system is stable if and only if

$$\mu_1 - \lambda t > 0 \quad \text{and} \quad \mu_2 - \lambda(1 - t + F(t)) > 0 \quad \text{for some } t \in [0,1], \tag{8}$$

where $F : [0,1] \to [0,1]$ is defined as $F(t) = \int_0^t f(x)dx$, denoting the proportion of patients who originally join the telemedicine service but end up needing in-person care. Following elementary algebraic rearrangements, the conditions in Equation (8) are equivalent to

$$\mu_2 > \lambda - \mu_1 + \lambda F(\mu_1/\lambda) \quad \text{if } \mu_1 \leq \lambda. \tag{9}$$

Assuming that Equation (9) holds, we say that the system is in mixed strategy Nash equilibrium under patient strategy $g^* : [0,1] \to [0,1]$ if and only if there exists $t^* \in [0,1]$ such that

$$g^*(t) = \begin{cases} 1 & \text{if } t \leq t^* \\ 0 & \text{o.w. ,} \end{cases} \tag{10}$$

and for all $t \in [0,1]$,

$$t^* = \arg\min_p \mathbb{E}\left[\mathbf{1}_{Tele}(p)\left(W_1(g^*_{-U}, p) + \mathbf{1}_{Follow-up}(U) W_2(g^*_{-U}, p)\right) + \left(1 - \mathbf{1}_{Tele}(p)\right) W_2(g^*_{-U}, p) \;\middle|\; U = t\right]. \tag{11}$$

The next proposition establishes the existence of a unique mixed strategy Nash equilibrium when patients' information granularity is refined.

Proposition 2 *Suppose that Equation* (9) *holds. In the refined information granularity regime, there exists a unique mixed strategy Nash equilibrium that satisfies conditions* (10) *and* (11)*, with* t^* *characterized as follows:*

1. *When* $\lambda + \mu_1 \leq \mu_2$*, all patients select the in-person visit, i.e.,* $t^* = 0$*;*
2. *Otherwise, the threshold* t^* *satisfies* $0 < t^* < 1$ *and is the unique solution to*

$$\mu_2 - (1 - t + F(t))\lambda = (1 - f(t))(\mu_1 - \lambda t).$$

Based on the characterization of t^* in Proposition 2, it can be shown that the refined patient equilibrium t^* is increasing in μ_1 and decreasing in μ_2, as formalized in Lemma B.1 in Appendix B.4. Furthermore, Equation (9) delineates the condition under which both queues maintain stability under some threshold type of patient strategy in the refined information regime. Proposition 2 further guarantees the attainment of a unique game-theoretic equilibrium under this strategy. Consequently, the following corollary directly follows, indicating that $\mathcal{S}^{refined}$ aligns with the parameter space characterized by Equation (9).

Corollary 2 *It holds that* $\mathcal{S}^{refined} = \{(\lambda, \mu_1, \mu_2, f) : \mu_2 > \lambda - \mu_1 + \lambda F(\mu_1/\lambda) \;\; \text{if } \mu_1 \leq \lambda\}$.

4. Impact of Information Granularity

In this section, we assess whether providing more information to patients, e.g., by employing predictive analytics in the online triage tool, can improve the system performance. In particular, we compare the patients' equilibria in the crude and refined information granularity regimes in terms of system stability and average patient waiting time.

4.1. Impact of Information Granularity on System Stability

In terms of system stability, it is consistently beneficial to provide each patient with more granular information about their likelihood of needing in-person follow-up after telemedicine. As established by Proposition 3 below, the refined information regime, in comparison to the crude information regime, possesses a strictly larger set of parameters over which the system admits a mixed strategy equilibrium with stable queues.

Proposition 3 *We have $\mathcal{S}^{crude} \subset \mathcal{S}^{refined}$.*

Figure 2 provides an illustration of Proposition 3, where we plot the parameter space in two dimensions with respect to μ_1 and μ_2. We set $\lambda = 0.5$ and $f(t) = t$ (which leads to $\mathbb{E}[X] = 0.5$). In the figure, the yellow region corresponds to $\mathcal{S}^{crude}$, with boundaries parameterized by (μ_1, μ_2) subject to $\mu_2 = \lambda - \mu_1 + \mu_1\mathbb{E}[X]$ and $\mu_2 = \lambda\mathbb{E}[X]$. For each combination of μ_1 and μ_2 in the yellow region, there exists a unique patient equilibrium with stable queues in the crude information regime. In comparison, the gridded region corresponds to $\mathcal{S}^{refined}$, whose boundaries are parameterized by (μ_1, μ_2) subject to $\mu_2 = \lambda - \mu_1 + \lambda F(\mu_1/\lambda)$ and $\mu_2 = \lambda\mathbb{E}[X]$. For each pair of μ_1 and μ_2 in the gridded area, the system possesses a unique patient equilibrium with stable queues in the refined information regime. As established in Proposition 3, the gridded area is strictly larger than the yellow area: When $\mu_1 < \lambda$, the boundary of the yellow region is strictly above that of the gridded region, as $\lambda - \mu_1 + \mu_1\mathbb{E}[X] > \lambda - \mu_1 + \lambda F(\mu_1/\lambda)$; when $\mu_1 \geq \lambda$, the boundaries of the yellow and gridded areas coincide with each other.

4.2. Impact of Information Granularity on Waiting Time

In Section 4.1 above, we have established that having patients make better-informed self-interested choice between care modalities leads to an expanded stability region for the system. In this section, we further compare the average steady-state waiting times under the patient crude and refined equilibria for fixed model parameters.

To avoid trivial comparison, we focus on the parameter region where the system is stable under some mixed strategy patient equilibrium for both the crude and refined information regimes,

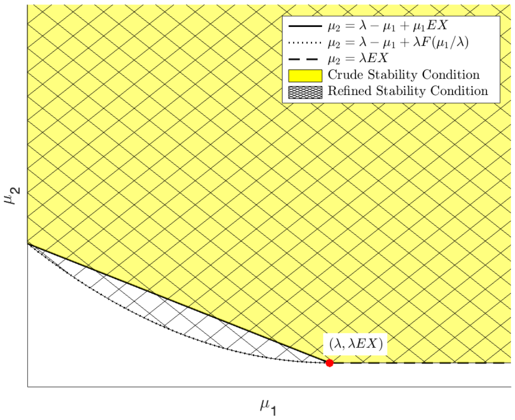


Figure 2 **Comparison Between the Refined and Crude Stability Conditions**

namely, we focus on the model parameters in $\mathcal{S}^{crude} \cap \mathcal{S}^{refined}$, for $\mathcal{S}^{crude}$ characterized in Corollary 1 and $\mathcal{S}^{refined}$ characterized in Corollary 2. We denote $h^c(p^*)$ and $h^r(t^*)$ as the average steady-state waiting times in equilibrium under the crude and the refined information regimes, respectively.

Theorem 1 *Fix λ and f. For any μ_1, there exists $\tilde{\mu}_2(\mu_1)$ such that the following holds:*

1. *When $\mu_2 \geq \mu_1 + \lambda$, the equilibrium strategies and expected steady-state waiting times satisfy*

$$p^* = t^* = 0 \quad \textit{and} \quad h^c(p^*) = h^r(t^*);$$

2. *When $\mu_1 + \lambda > \mu_2 \geq \tilde{\mu}_2(\mu_1)$, the expected steady-state waiting times satisfy*

$$h^c(p^*) \geq h^r(t^*),$$

and there exists $\bar{\mu}_2(\mu_1) \in (\tilde{\mu}_2(\mu_1), \mu_1 + \lambda)$ such that the equilibrium strategies satisfy

$$(a)\ 0 \leq p^* < t^* \textit{ for } \mu_2 > \bar{\mu}_2(\mu_1), \quad \textit{and} \quad (b)\ p^* \geq t^* > 0 \textit{ for } \mu_2 \leq \bar{\mu}_2(\mu_1);$$

3. *When $\tilde{\mu}_2(\mu_1) > \mu_2 > \lambda\mathbb{E}[X]$, there exist equilibrium strategies and expected waiting times that satisfy*

$$p^* > t^* \quad \textit{and} \quad h^c(p^*) < h^r(t^*).$$

The implicit expressions for $\tilde{\mu}_2(\mu_1)$ and $\bar{\mu}_2(\mu_1)$ in Theorem 1 are provided in Appendix B.6. In addition, we emphasize that in Case 3 of Theorem 1, while we only formally prove the existence of such orderings between patient strategies and average steady-state waiting times (i.e., $p^* > t^*$ and $h^c(p^*) < h^r(t^*)$), we conjecture that such orderings hold across the entire parameter region for Case 3, specifically when $\tilde{\mu}_2(\mu_1) > \mu_2 > \lambda\mathbb{E}[X]$. Unfortunately, due to limited analytical tractability,

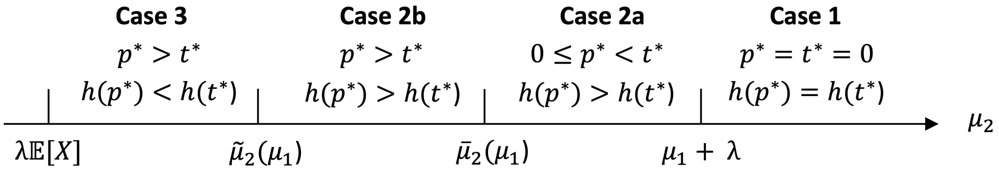


Figure 3 **Average Waiting Time Comparison Under Refined and Crude Equilibria for Fixed μ_1**

we are unable to provide a formal proof for this conjecture. Nevertheless, extensive numerical experiments consistently support the same conclusion.

Figure 3 provides a visualization of the four regions characterized in Theorem 1 as the value of μ_2 varies. In particular, Case 1 corresponds to the parameter regime where the in-person service rate is sufficiently high ($\mu_2 > \mu_1 + \lambda$) so that all patients opt for in-person services in both the crude and refined information granularity regimes ($p^* = t^* = 0$). As a result, the expected waiting times in equilibrium are equivalent in both information regimes ($h^c(p^*) = h^r(t^*)$). Compared to Case 1, Cases 2 and 3 have relatively low in-person service rates, resulting in non-negligible proportions of patients who select telemedicine services. In Case 2a, where the in-person service rate falls below $\mu_1 + \lambda$ but remains above $\bar{\mu}_2(\mu_1)$, the equilibrium in the refined information regime induces more patients (with a total proportion equal to t^*) to select telemedicine. The refined information regime also results in strictly shorter expected steady-state waiting time in equilibrium than the crude information regime. In Cases 2b and 3, the in-person service rate is lower than $\bar{\mu}_2(\mu_1)$, and, unlike Case 2a, fewer patients select telemedicine in the refined information equilibrium than in the crude information equilibrium ($p^* > t^*$). Furthermore, while the expected steady-state waiting time in equilibrium is shorter in the refined information regime in Case 2, Case 3 exhibits alternative orderings of the expected waiting times in the crude and refined information regimes. Specifically, Case 3 has patients experience less waiting on average when the available information is crude ($h^c(p^*) < h^r(t^*)$). Notably, Case 3 characterizes a counterintuitive phenomenon in which providing patients with more information about their likelihood of needing in-person follow-ups after telemedicine "backfires" and increases the expected steady-state waiting time in equilibrium. In other words, under the refined information equilibrium in Case 3, although patients make better-informed self-interested decisions in choosing care modalities, the overall system performance (in terms of the average waiting time) turns out to be worse.

For any fixed μ_1, Theorem 1 above compares the choices made by patients regarding care modalities and the expected waiting times in equilibrium across four regions delineated by $\lambda\mathbb{E}[X]$, $\tilde{\mu}_2(\mu_1)$, $\bar{\mu}_2(\mu_1)$, and $\mu_1 + \lambda$. In complement to Theorem 1, Proposition 4 further asserts the continuity of the mappings $\tilde{\mu}_2(\mu_1)$ and $\bar{\mu}_2(\mu_1)$ with respect to μ_1.

Proposition 4 *Let $\bar{\mu}_2(\cdot)$ and $\tilde{\mu}_2(\cdot)$ be as defined in Theorem 1. It holds that $\bar{\mu}_2(\cdot)$ and $\tilde{\mu}_2(\cdot)$ are continuous.*

It follows from Proposition 4 that the two-dimensional parameter space with respect to μ_1 and μ_2 can be partitioned into four regions in analogue to those characterized in Theorem 1. In each region of the two-dimensional parameter space, the orderings of patient choices (p^* and t^*) and expected equilibrium waiting time ($h^c(p^*)$ and $h^r(t^*)$) stay the same as those characterized in Theorem 1. We demonstrate such partition in Figure 4, on a (μ_1, μ_2) grid, when $\lambda = 0.5$, $f(t) = t^{1/5}$, and $\mathbb{E}[X] = 5/6$. Note that the parameter space is divided into four regions colored in white, green, blue, and red. The white region at the left bottom denotes the unstable parameter regime (outsides $\mathcal{S}^{crude} \cap \mathcal{S}^{refined}$) which we do not consider for comparison. For the other regions, there exist unique patient equilibria with stable queues in both the crude and refined information regimes, which enables non-trivial comparison of the patients equilibrium strategies and the corresponding average steady-state waiting times. In particular, the green region corresponds to Case 1 in Theorem 1, where all patients opt in for the in-person service in both information granularity regimes, thus leading to equivalent average waiting times. The blue region corresponds to Case 2 with $h^c(p^*) \geq h^r(t^*)$ in Theorem 1. Lastly, the red region corresponds to Case 3 with $h^c(p^*) < h^r(t^*)$ in Theorem 1, which exhibits the counterintuitive phenomenon that providing patients with more information leads to increased average waiting time in equilibrium. Intuitively, we observe that in the red region, the telemedicine service rate is much higher than the in-person service rate. In the crude information regime, p^* tends to approach or equal 1. However, upon receiving additional information, some patients reconsider and opt for in-person visits, leading to $t^* < p^*$. This influx of patients seeking in-person visits subsequently elevates the average waiting time in the in-person queue. Consequently, this also contributes to the waiting time for patients requiring in-person follow-ups after telemedicine. As a result, the overall average waiting time of the system surpasses that in the crude information regime, exemplifying the "tragedy of the commons."

5. System's First Best Solution

In the system's first best problem, the system manager makes a centralized routing decision for each patient, aiming to minimize the average total waiting time in the system. In terms of information availability, we assume that the system manager knows the realization of each patient's specific type and the corresponding likelihood of requiring an in-person follow-up visit subsequent to a telemedicine appointment. It is important to note that this assumption is without loss of generality, as having more detailed information about patients' types and follow-up probabilities leads to equivalent or strictly improved centralized decision-making.

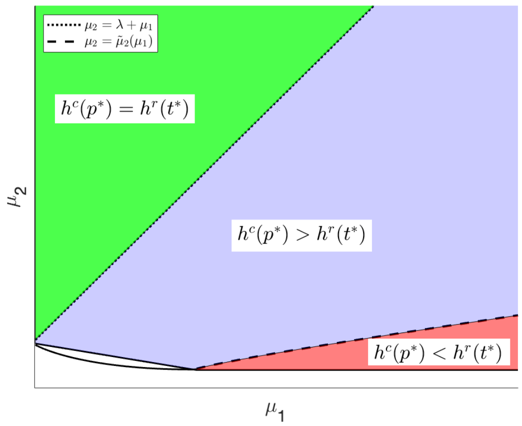


Figure 4 **Average Waiting Time Comparison Under Refined and Crude Equilibria**

Recall from Equation (1) that given patients' strategies g, the steady-state total waiting time for patient of type $U = t$ is given by $W(g,t)$, which is a random variable. The average steady-state total waiting time in the system is $\mathbb{E}[W(g,U)]$, where the expectation is taken with respect to both queueing dynamics and random patient type U. The system's first best problem is given by

$$\min_{g} \ \mathbb{E}[W(g,t)]. \tag{12}$$

For the system's first best problem (12), we show that similar to the refined equilibrium, the system's first best solution is also a threshold-type solution; see the formal statement in Proposition 5 below. That is, it is without loss of optimality to restrict to the family of threshold policies under which the system manager routes patients with type values below a certain threshold to telemedicine consultations and routes the rest of the patients to in-person service. Solving the first best problem (12) is then equivalent to finding the optimal threshold. Rewriting problem (12) by restricting to threshold strategies, we get

$$\min_{t\in[0,1]} \ h^{fb}(t) := \frac{t}{\mu_1 - \lambda t} + \frac{1-t+F(t)}{\mu_2 - \lambda(1-t+F(t))}, \tag{13}$$

where $h^{fb}(t)$ denotes the first best average steady-state waiting time across all patients under a threshold-type routing policy with threshold t.

Because both the patients' equilibrium strategy in the refined information granularity regime and system's first best solution are threshold-type, it follows from the analysis in Section 3.2 that the system stability condition under centralized routing is identical to that in the refined information granularity regime, as summarized in Corollary 3.

Corollary 3 *There exists some centralized routing strategy g under which the system is stable if and only if $(\lambda,\mu_1,\mu_2,f)\in\mathcal{S}^{refined}$.*

Recall from Section 3.2 that $F(t)=\int_0^t f(x)dx$, which denotes the proportion of patients who originally join the telemedicine service but end up needing an in-person follow-up visit under the t-threshold strategy. Based on this definition, Proposition 5 states the optimal solution to the system's first best problem.

Proposition 5 *For $(\lambda,\mu_1,\mu_2,f)\in\mathcal{S}^{refined}$, the system's first best solution is threshold-type, with a unique optimal threshold $\bar{t}$ such that patients with $t<\bar{t}$ are routed to the telemedicine queue and patients with $t\geq\bar{t}$ are routed to the in-person queue. In addition, $\bar{t}$ satisfies:*

1. *When $\lambda+\sqrt{\mu_1\mu_2}\leq\mu_2$, $\bar{t}=0$, i.e., routing all patients to the in-person queue;*
2. *When $\lambda+\sqrt{\mu_1\mu_2}>\mu_2$, $0<\bar{t}<1$, and $\bar{t}$ is the unique solution to*

$$\mu_1(\mu_2-(1-t+F(t))\lambda)^2=\mu_2(\mu_1-\lambda t)^2(1-f(t)).$$

5.1. Comparison of System's First Best Solution and Patients' Equilibrium

In this section, we compare the patients' equilibrium in the refined information granularity regime to the system's first best solution.

As the equilibrium strategy of the patients and the optimal centralized solution of the system both adopt threshold-based policies, our initial comparison involves assessing the thresholds t^* and $\bar{t}$. Recall that these thresholds characterize, respectively, the strategy adopted by patients in equilibrium and the system's first best solution.

To facilitate the comparison, we introduce the notation $\hat{t}:=f^{-1}(1-\mu_2/\mu_1)$. In essence, $\hat{t}$ represents the patient type who is indifferent in selecting either service when no other patients are present in the system, namely, $1/\mu_1+f(\hat{t})/\mu_2=1/\mu_2$. With the aid of $\hat{t}$, we establish criteria for the two thresholds in Proposition 6 below.

Proposition 6 *The following comparison holds for t^* and $\bar{t}$:*

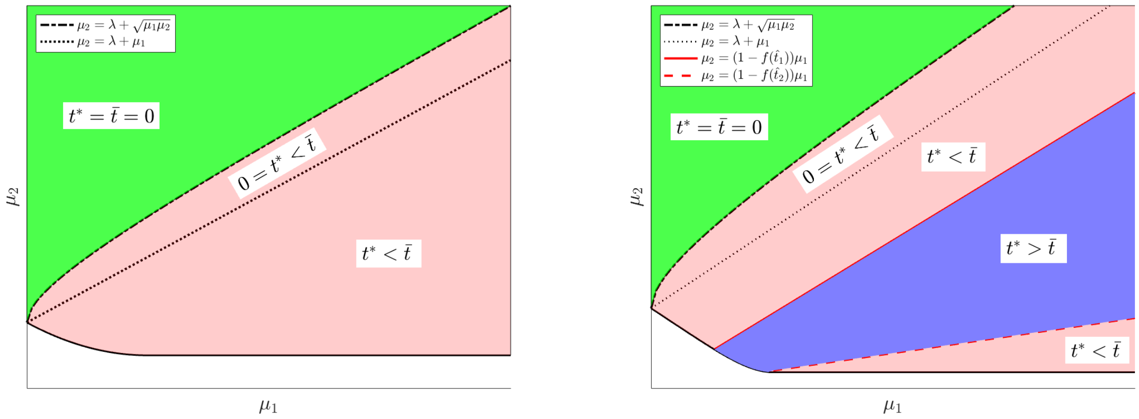
1. *When $\mu_1\leq\mu_2$, $t^*\leq\bar{t}$;*
2. *When $\mu_1>\mu_2$,*
   (a) *If $1-\hat{t}+F(\hat{t})-(1-f(\hat{t}))\hat{t}<0$, $t^*>\bar{t}$;*
   (b) *If $1-\hat{t}+F(\hat{t})-(1-f(\hat{t}))\hat{t}=0$, $t^*=\bar{t}$;*
   (c) *If $1-\hat{t}+F(\hat{t})-(1-f(\hat{t}))\hat{t}>0$, $t^*<\bar{t}$.*

Proposition 6 immediately implies that if the following sufficient condition holds, there are fewer patients opting for the telemedicine service than the number desired by the system's first best solution, i.e., $t^* \leq \bar{t}$:

$$1 - t + F(t) - (1 - f(t))t \geq 0 \quad \forall t \in [0,1]. \tag{C}$$

To provide some intuitive interpretation of condition (C), we note that the functional forms of f that meet this condition typically possess a higher average probability of needing in-person follow-ups. For instance, the function $f(t) = t^2$ satisfies condition (C), resulting in an average follow-up probability equal to $\mathbb{E}[U^2] = 1/3$. Moreover, any function f where $f(t) \geq t^2$ for all $t \in [0,1]$ yields a higher average follow-up probability than $\mathbb{E}[U^2]$ and therefore fulfills the condition. Conversely, a function such as $f(t) = t^4$ leads to a relatively lower expected follow-up probability equal to $\mathbb{E}[U^4] = 1/5$, thereby violating this condition. It follows from the intuition behind condition C that when the average follow-up probability is relatively high, patients tend to exhibit more risk aversion in their individual decisions regarding telemedicine usage. Consequently, a smaller portion of patients utilizes telemedicine in the patients' equilibrium compared to the system's first best, i.e., $t^* \leq \bar{t}$. In contrast, when the average follow-up probability is relatively low, patients lean towards being more risk-seeking in their preference for telemedicine. As a consequence, the proportion of patients using telemedicine in the patients' equilibrium surpasses the optimal level observed in the system's first best, namely, $t^* > \bar{t}$.

Figure 5 illustrates Proposition 6, depicting the comparison between t^* and $\bar{t}$ in the two-dimensional parameter space defined by (μ_1, μ_2). In particular, Figure 5(a) and Figure 5(b) correspond to two distinct functions f. Figure 5(a) represents the scenario where f satisfies condition (C), while Figure 5(b) illustrates a situation where condition (C) is violated. For Figure 5(a), the parameters are set as $\lambda = 0.5$, $f(t) = t$, $\mu_1 \in [0,2]$, and $\mu_2 \in [0, 2.915]$. Meanwhile, for Figure 5(b), the parameters are specified as $\lambda = 0.5$, $f(t) = t^4$, $\mu_1 \in [0,2]$, and $\mu_2 \in [0, 2.396]$. We make the following observations. First, the white regions in Figure 5 signify system instability. In the remaining parameter space, when $\mu_1 \leq \mu_2$, we have $t^* \leq \bar{t}$, which corresponds to Case 1 in Proposition 6. Otherwise, the relationship between t^* and $\bar{t}$ depends on the properties of f. In Figure 5(a), where f satisfies condition (C), we consistently have $t^* \leq \bar{t}$. In contrast, Figure 5(b) demonstrates a scenario where f satisfies the assumption of Case 2(a) in Proposition 6. Consequently, a continuous blue region emerges where $t^* > \bar{t}$. It is noteworthy that the presence of the blue region in Figure 5(b) is not incidental. For any general functional form of f, it can be derived that the region where $t^* > \bar{t}$ is consistently bounded by two linear lines expressed as $\mu_2(\mu_1) := (1 - f(v_1))\mu_1$ and $\mu_2(\mu_1) := (1 - f(v_2))\mu_1$. Here, v_1 and v_2 represent the two distinct roots of $1 - v + F(v) - (1 - f(v))v = 0$.



(a) f satisfies condition C

(b) f violates condition C

Figure 5 **Comparison of t^* and $\bar{t}$ in the Refined Information Equilibrium and System's First Best**

6. Coordination Mechanism by Priority Rules

In this section, we focus on the parameter regime where the system's first best solution strictly outperforms the patients' equilibrium in the refined information granularity regime, i.e., $\mu_2 < \lambda + \sqrt{\mu_1\mu_2}$. We explore a family of priority rules as a potential mechanism to induce waiting-time incentives for patients' choices, aiming to guide the system towards its first best solution. More specifically, in the sub-region where the patients' equilibrium in the refined information granularity regime results in longer average waiting times compared to that in the crude information granularity regime, our objective is to implement the mechanism to transform the information disadvantage into an advantage. Additionally, within the rest of the parameter regime, where providing patients with more information leads to shorter average waiting times than the waiting times in the equilibrium without information provision, we leverage the mechanism to further minimize or close the performance gap and move towards the system's first best solution.

To design the priority mechanism, we observe from Proposition 6 in Section 5.1 that there exist two possible orderings between the thresholds of selecting telemedicine in the patients' equilibrium and the system's optimal solution. Depending on the model primitives, the threshold in the patients' equilibrium can be higher or lower than that in the system's optimal solution. Given these two possible scenarios, the design of the priority rule aims to encourage patients to choose the modality that is more desired under the first best solution by reducing the overall waiting time in the system for those patients.

To formalize this idea, we consider a family of priority rules that are characterized by a probability parameter $0 \leq q \leq 1$, where q is a control variable to be optimized. Although parameterized by a single parameter q, the direction of assigning priority is contingent upon the relationship between t^* and $\bar{t}$. Specifically, if the number of patients choosing telemedicine service is fewer than that desired by the system's first-best solution (i.e., $t^* < \bar{t}$), priority is allocated to in-person follow-up visits to encourage patients to choose telemedicine. In particular, patients requiring an in-person follow-up visit post telemedicine are prioritized with probability q, i.e., moved to the front of the in-person queue, independently and identically (i.i.d). Conversely, if more patients choose telemedicine service than in the first-best solution (i.e., $t^* > \bar{t}$), priority is given to initial in-person visits to encourage patients to opt for the in-person modality. That is, patients who initially choose in-person visits are prioritized with probability q in an i.i.d manner. By construction, we assign no priority when $q = 0$ and full priority when $q = 1$. Moreover, assigning a value $0 < q < 1$ leads to partial prioritization, where, on average, a proportion q of the eligible patient population receives priority. The endogeneity of q enables us to offer varying levels of waiting-time incentives, contingent upon the difference between t^* and $\bar{t}$.

We explore the combined impact of information provision and prioritization. Specifically, in the refined information granularity regime, we formally show in the proof of Theorem 2 below that under any priority rule parameterized by $q \in [0,1]$, the system induces threshold-type patient equilibria. Next, we consider all possible patient equilibria achieved by the family of priority rules with $q \in [0,1]$ and identify the equilibrium with the shortest average waiting time. We denote this shortest average waiting time by $h^p(q^*)$, where q^* represents the parameter of the corresponding priority rule. We then say that the priority rule with parameter q^* is able to induce the system's first best solution if $h^p(q^*)$ is the same as $h^{fb}(\bar{t})$, i.e., the optimal average waiting time achievable under centralized control.

Theorem 2 below characterizes conditions under which the priority rule can induce the system's first best solution. To state the results, let $S := \mu_2 - \lambda(1 - \bar{t} + F(\bar{t})) - (1 - f(\bar{t}))(\mu_1 - \lambda\bar{t})$, and $M := (1 - \bar{t} + F(\bar{t}))((1 - \bar{t})f(\bar{t}) + F(\bar{t}))(\mu_1 - \lambda\bar{t})$.

Theorem 2 *Assume that $(\lambda, \mu_1, \mu_2, f) \in \mathcal{S}^{refined}$. The following holds for the priority rules:*

1. *If $\mu_2 \geq \mu_1 + \lambda$, applying priority rules cannot further reduce the average waiting time in the refined information equilibrium, i.e., $h^p(q^*) = h^r(t^*) > h^{fb}(\bar{t})$;*
2. *If $\mu_2 < \mu_1 + \lambda$,*
   (a) *When $t^* < \bar{t}$: If $\lambda^2 M \geq \mu_2(\mu_2 - \lambda F(\bar{t}))S$, we have*

$$q^* = \frac{\mu_2^2 S}{\lambda(\mu_2 F(\bar{t})S + \lambda M)},$$

and the priority rule with parameter q^* (which prioritizes in-person follow-ups) induces the system's first best solution, i.e., $h^r(t^*) > h^p(q^*) = h^{fb}(\bar{t})$; otherwise, we have $q^* = 1$, and the priority rule with parameter q^* does not induce the system's first best solution, i.e., $h^r(t^*) > h^p(1) > h^{fb}(\bar{t})$;

(b) *When $t^* > \bar{t}$: If $\lambda^2 M \geq \mu_2(\lambda(1-\bar{t}) - \mu_2)S$, we have*

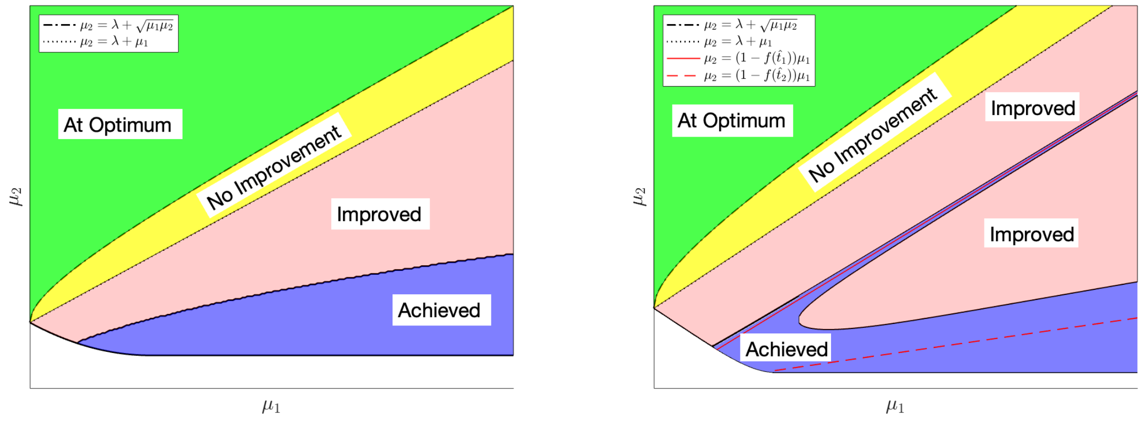
$$q^* = \frac{\mu_2^2 S}{\lambda\left(\mu_2(1-\bar{t})S - \lambda M\right)},$$

and the priority rule with parameter q^* (which prioritizes initial in-person visits) induces the system's first best solution, i.e., $h^r(t^*) > h^p(q^*) = h^{fb}(\bar{t})$; otherwise, we have $q^* = 1$, and the priority rule with parameter q^* does not induce the system's first best solution, i.e., $h^r(t^*) > h^p(1) > h^{fb}(\bar{t})$.

In light of Theorem 2, the priority rule successfully induces the system's first best solution if the conditions in Cases 2(a) and 2(b) are satisfied. In cases where these conditions are not fulfilled, adopting full priority continues to result in reduction in the average waiting time. The only exception to this improvement is observed when $\mu_2 \geq \mu_1 + \lambda$, where the priority rule does not impact the average waiting time.

Figure 6 illustrates Theorem 2, with the same numerical configurations as those in Figure 5. In Figure 6, the blue region denotes the parameter combinations (μ_1, μ_2) where using the priority rule induces the system's first best solution. The red region marks the parameter combinations where applying the priority rule strictly reduces the average patient waiting time in the refined information equilibrium, but falls short of achieving the system's first best solution. The yellow region depicts the parameter region where the priority rule does not have any impact on improving the average waiting time. Moreover, it is noteworthy that for the numerical instances considered here, the red region in Figure 4 is contained in the blue region in Figure 6. This observation implies that in the parameter region where the average waiting time in the refined information equilibrium exceeds that in the crude information equilibrium, we can turn the information disadvantage into an advantage by simultaneously applying the priority rule. It is crucial to emphasize that, while we cannot formally prove this statement for general model primitives due to limited analytical tractability, extensive numerical experiments consistently support the same conclusion.

Finally, in establishing Theorem 2, we first identify the shortest average waiting time $h^p(q^*)$ among all potential equilibria under priority rules, then retrieve the corresponding priority parameter q^*. Nevertheless, it remains unclear a priori whether the patient equilibrium with an average



(a) f satisfies condition C

(b) f violates condition C

Figure 6 **Conditions Under Which the Priority Rule Induces the First Best Solution**

waiting time of $h^p(q^*)$ is unique under the priority rule with the specified parameter q^*. Indeed, the uniqueness of patient equilibrium under priority rules is highly nontrivial and may not hold across all values of the parameter q. To mitigate concerns about encountering multiple equilibria under the priority rule, we provide conditions in Proposition 7 to verify the uniqueness of the equilibrium under any given priority rule. To state the results, we define a constant $\bar{q} \in \mathbb{R}$ and a function $\hat{q} : [0,1] \to \mathbb{R}$, whose explicit expressions are provided in Appendix B.10.

Proposition 7 *Assume that $(\lambda, \mu_1, \mu_2, f) \in \mathcal{S}^{refined}$. Under the priority rule with parameter q, a unique threshold-type Nash equilibrium exists if the following conditions hold:*

1. *When $t^* < \bar{t}$, at least one of the following three conditions is satisfied: $\mu_1 \le \lambda$, $\mu_2 \ge \mu_1 - (1 - \mathbb{E}[X])\lambda$, or $q < \bar{q}$;*
2. *When $t^* > \bar{t}$, $0 \le q < \hat{q}(t)$ for all $t < t^*$.*

7. Case Study

In this section, we conduct a comprehensive case study using real-world outpatient data from a large academic hospital in Maryland. The goal of the case study is to provide a step-wise demonstration of how to implement our theoretical framework in a real-world clinic setting and how to estimate performance improvement. The case study consists of the following key components:

1. **Prediction model:** In our theoretical model, we assume that the distribution for the in-person follow-up probability after telemedicine consultations is given. To facilitate real-world implementation, we construct a logistic regression model using visit-level data to predict whether each patient will necessitate an in-person follow-up after receiving treatment via

telemedicine. We illustrate how the developed prediction model can be integrated into our theoretical framework and applied during real-world online triage.

2. **Model calibration:** We demonstrate how to calibrate the model parameters using real-world data. In addition to the arrival and service rates (λ, μ_1, μ_2), we illustrate how the distribution of the follow-up probability X (alternatively, the function f) can be retrieved from the constructed logistic regression model.
3. **Implementation of proposed policies:** Utilizing the prediction model and calibrated parameters, we identify the parameter region to which the hospital operations belong, which determines the directional impact of information provision and priority rules. In a more quantitative analysis, we calculate the average patient waiting time under the proposed policies and quantify the improvement over the benchmark scenario.

For the rest of this section, we review the data set in Section 7.1 and develop the prediction model in Section 7.2. Then in Section 7.3, we calibrate the model and compare the system performance under our proposed policies to the benchmark scenario.

7.1. Data Set

We obtain our research data for a large academic hospital in Maryland from the Maryland Health Services Cost Review Commission. The focal hospital started providing telemedicine options for preprocedural assessment on January 1, 2021. To analyze the efficacy of telemedicine treatment in substituting for in-person preprocedural assessment, we examine the data related to preprocedural assessment from January 1, 2021, to September 30, 2023, identified through the primary diagnosis code Z01.818 within the ICD-10 system. The data set contains a total of 3,275 visits, among which 210 (6.412%) were conducted through telemedicine. These visits involve a total of 3,159 unique patients, among whom 209 unique patients had telemedicine service.

The data contain both the visit-level clinical information and patient-level demographic characteristics. Unique patient identifiers are provided to track patients' visits, follow-ups, and diagnoses over time. At the encounter level, the data record the time of the visit, care modality (i.e., in-person or telemedicine), patient demographic information, insurance plan, patient comorbidities (i.e., the Charlson comorbidity index), arrival source (i.e., home or another care site), and up to six additional diagnosis codes that specify the clinical focus of preprocedural assessment, e.g., E00–E89 (nutrition), H00–H59 (eyes), and K00–K95 (digestion). The definition and description of each diagnosis code can be found on the CDC website.

Table A.1 in Appendix A provides summary statistics for the data fields. Additionally, we calculate the number of days from the focal preprocedural assessment until an in-person follow-up visit (if any) and report the mean and standard deviation. Moreover, we include a measure to signal the overall care needs of each patient, defined as the total number of outpatient visits for all purposes conducted by this patient during 2020 (the year before the start of our focal data set). From Table A.1, we see that the patient population who received preprocedural assessment exhibits sufficient heterogeneity in terms of patient characteristics.

7.2. Prediction Model

The objective of the prediction is to forecast the likelihood of an in-person follow-up after the focal visit. To this end, we identify whether each visit is followed by another visit within 7, 14, 21, and 30 days. These indicators, labeled as *followup7*, *followup14*, *followup21*, and *followup30*, are our target variables for prediction. To evaluate the performance of the regression model, we partition the data set into training and test sets. The training set comprises 90% (2,947 records) of the data, while the test set includes the remaining 10% (328 records).

We construct a logistic regression model by incorporating various sets of predictors. The first set contains patient-level details such as age, sex, ethnicity, race, insurance, arrival source, Charlson comorbidity index, and the total number of outpatient visits by each patient in 2020. In addition, for each of the 22 diagnosis codes, we introduce a binary variable indicating the presence of that specific diagnosis code during the focal visit. Moreover, we incorporate time-fixed effects using indicators for the year and quarter. Lastly, we include a binary indicator for the visit modality, namely, telemedicine or in-person. Results of the logistic regression are partially presented in Table 1 for discussion, with the rest of the results relegated to Appendix A.

We draw several interesting observations based on the results presented in Table 1. First, the indicator for care modality is statistically significant and associated with a positive coefficient. This suggests that telemedicine visits are positively correlated with the need for in-person follow-up, which validates our modeling assumption. Second, distinct insurance types exhibit varying levels of association with in-person follow-up needs. For example, with everything else held constant, self-pay patients are less inclined to pursue in-person follow-ups compared to those with commercial insurance. Conversely, patients covered by Medicare or Medicaid plans are more likely to seek in-person follow-ups than those utilizing commercial insurance. Third, the arrival source of the focal visit, whether from home or from other care sites, emerges as a significant predictor of in-person follow-up needs. Specifically, patients arriving from other care sites have a higher likelihood of requiring in-person follow-ups than those arriving from home. Fourth, the total number of outpatient visits in 2020, created to signal the patient's overall care needs, is positively associated with the likelihood of experiencing in-person follow-up visits. Lastly, the binary indicators for various

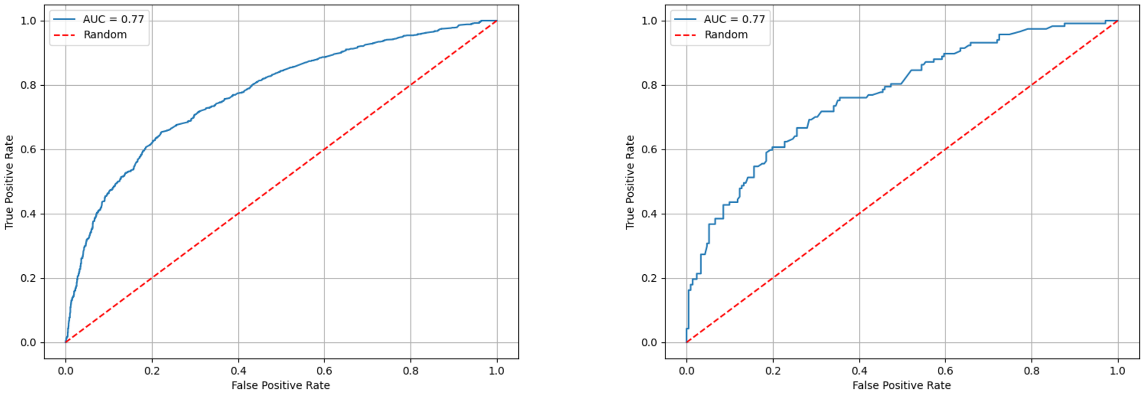
Table 1 Regression Results (Part 1)

	followup7	followup14	followup21	followup30
	(1)	(2)	(3)	(4)
Tele-Visit	0.475^{**}	2.231^{***}	3.110^{***}	3.326^{***}
	(0.235)	(0.360)	(0.539)	(0.612)
Age Group (20-39)	0.078	0.092	0.056	-0.013
	(0.187)	(0.187)	(0.185)	(0.183)
Age Group (40-59)	-0.115	-0.041	-0.080	-0.099
	(0.170)	(0.169)	(0.168)	(0.165)
Age Group (60-79)	-0.411^{**}	-0.277	-0.244	-0.268
	(0.186)	(0.185)	(0.182)	(0.179)
Age Group (80+)	0.356	0.595^{**}	0.579^{**}	0.514^{*}
	(0.269)	(0.269)	(0.267)	(0.265)
Sex (Male)	-0.652^{***}	-0.602^{***}	-0.574^{***}	-0.551^{***}
	(0.096)	(0.094)	(0.093)	(0.091)
Ethnicity	-0.137	-0.050	-0.033	-0.038
(Not Hispanic or Latinx)	(0.221)	(0.225)	(0.225)	(0.223)
Ethnicity	-23.784	-23.761	-23.805	-3.107^{***}
(Unkown)	(19377.296)	(19454.466)	(19471.075)	(1.052)
Insurance (Medicaid)	0.537^{***}	0.631^{***}	0.713^{***}	0.798^{***}
	(0.164)	(0.163)	(0.161)	(0.160)
Insurance (Medicare)	0.125	0.115	0.164	0.255^{*}
	(0.136)	(0.138)	(0.136)	(0.134)
Insurance (Self pay)	-1.146^{***}	-0.940^{**}	-0.939^{**}	-0.887^{**}
	(0.379)	(0.366)	(0.370)	(0.366)
Insurance (Other)	-0.028	-0.021	0.065	0.170
	(0.181)	(0.185)	(0.182)	(0.178)
Source of Arrival	0.532^{***}	0.476^{***}	0.442^{***}	0.370^{***}
(Other Care Sites)	(0.092)	(0.092)	(0.091)	(0.090)
Number of Visits in 2020	0.132^{***}	0.132^{***}	0.145^{***}	0.168^{***}
	(0.042)	(0.042)	(0.042)	(0.043)
Charlson Index	-0.170	-0.394^{**}	-0.225	-0.274
	(0.156)	(0.200)	(0.202)	(0.202)
Diagnosis Codes	-1.863^{**}	-1.507^{**}	-1.415^{**}	-1.253^{*}
(E00-E89, Nutrition)	(0.753)	(0.690)	(0.667)	(0.645)
Diagnosis Codes	1.596^{***}	1.608^{***}	1.547^{***}	1.549^{***}
(H00-H59, Eyes)	(0.277)	(0.278)	(0.278)	(0.278)
Diagnosis Codes	1.229^{***}	1.252^{***}	1.173^{***}	1.099^{***}
(K00-K95, Digestion)	(0.277)	(0.272)	(0.273)	(0.273)
Intercept	0.024	-0.456	-0.358	-0.202
	(0.341)	(0.356)	(0.361)	(0.360)
Observations	3275	3275	3275	3275

Note: *p<0.1; **p<0.05; ***p<0.01

diagnosis codes exhibit significant predictive power. As partially outlined in Table 1, patients with nutrition issues are less prone to in-person follow-ups, while those with eye or digestion issues are more likely to have in-person follow-ups, possibly driven by the necessity for physical examinations.

To evaluate the prediction accuracy of the logistic regression model, we analyze the corresponding ROC (receiver operating characteristic) curves on both the training and test data, as illustrated in Figure 7. Notably, the model attains an AUC (area under the ROC curve) of 0.77 for both the training and test sets. The consistent AUC values indicate effective and robust performance, alleviating concerns of overfitting.



(a) ROC curve on training data set

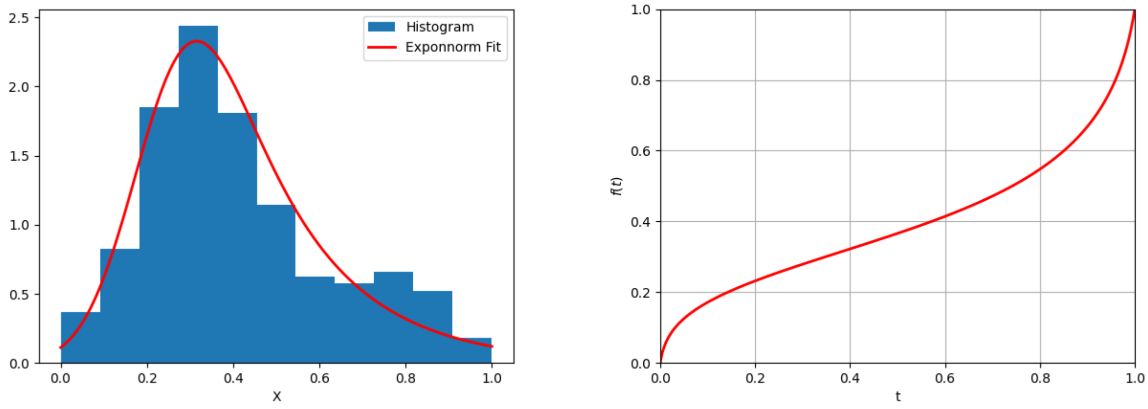
(b) ROC curve on test data set

Figure 7 **Performance of the Prediction Model**

7.3. Model Calibration

In this section, we calibrate the parameters for our theoretical model, i.e., $(\lambda, \mu_1, \mu_2, f)$, using the data set and the constructed logistic regression model.

In order to retrieve the functional form of f, we first estimate the distribution of in-person follow-up probabilities for our patient population. Based on the inverse-transform method, the function f is derived as the inverse of the cumulative distribution function representing these in-person follow-up probabilities. To this end, we assume that all the encounters in our data set occurred via telemedicine, and we focus on the seven-day follow-ups. Subsequently, we apply the logistic regression model to forecast the likelihood of each appointment requiring an in-person follow-up. Figure 8(a) demonstrates the histogram of these probabilities, which can be considered as the empirical distribution of the random variable X. We then use the Python package "Fitter" to identify the probability distribution that best aligns with the observed data. Based on the criterion of the smallest sum of squared errors, the most suitable distribution is the truncated exponential modified normal distribution (exponnorm) with support between 0 and 1. The parameters of the truncated exponnorm distribution are $K = 1.894$, $loc = 0.208$, and $scale = 0.108$. To visually demonstrate the goodness of fit, we depict the density function of the fitted exponnorm distribution via the red line in Figure 8(a). Subsequently, we retrieve the function f through the inverse cumulative distribution function of the fitted exponnorm distribution, as shown in Figure 8(b).



(a) Histogram with Exponnorm fit

(b) f by inverting Exponnorm CDF

Figure 8 **Fitted Distribution**

Having estimated the functional form of f, we proceed to calibrate the values of the arrival and service rates from the data. Assuming the hospital operates for 8 hours on weekdays, we calculate the hourly arrival rate for patients requiring procedural assessment as $\lambda = 3,275/(716 \times 8) = 0.572$, where 3,275 represents the total number of patient encounters over 716 weekdays in our data set. To determine a suitable telemedicine service rate μ_1, we refer to the findings of Meyer et al. (2023), who examine telemedicine duration for various patient types. According to Table 1 in Meyer et al. (2023), the average telemedicine service duration for return patients is 23 minutes, corresponding to a service rate of 2.6 encounters per hour. Considering that preprocedural assessment only constitutes a portion of all outpatient encounters, we further adjust the service rate based on the proportion of preprocedural assessment among all outpatient encounters. Specifically, we set $\mu_1 = 2.6 \times (3,275/219,734) = 0.039$, where 219,734 is the total number of outpatient visits at the hospital over the same period as our data. Finally, to estimate the in-person service rate μ_2, we scale μ_1 proportionally by the ratio between telemedicine throughput and in-person visit throughput. Consequently, we have $\mu_2 = \mu_1 \times (3,065/210) = 0.566$. We note that our theoretical model assumes single servers at both queues to gain analytical tractability. In this context, the calibrated values of μ_1 and μ_2 can be interpreted as the effective service rates of "super servers," accounting for potential variations in server counts in real-world scenarios.

7.4. Performance Evaluation

With the parameters $(\lambda, \mu_1, \mu_2, f)$ derived above, we apply the theoretical results to the hospital setting. First, we find that the hospital operates in the stable parameter region under both information regimes, i.e., $(\lambda, \mu_1, \mu_2, f) \in \mathcal{S}^{crude}$. Subsequently, we characterize the patient equilibrium in the two information granularity regimes, as well as the system's first best solution. Table 2 below lists the values of p^*, t^*, and $\bar{t}$, corresponding to the probability of selecting telemedicine service in the crude information equilibrium, the threshold for choosing telemedicine in the refined information equilibrium, and the threshold of the system's first best solution. We note that the telemedicine adoption rate observed in the data is 0.064, which is closer to $p^* = 0.043$ in the crude information equilibrium than $t^* = 0.039$ of the refined information equilibrium. This suggests that the current hospital operations resemble more closely to the crude information granularity regime, which is reasonable given that the hospital has not employed any online triage tool to provide information and guide patients' decisions. In Table 3, we further calculate the resulting average patient waiting times under both patient equilibria and the system's first best solution. In the crude information equilibrium, the average waiting time across all patients is approximately 3 weeks (14.48 days), with the waiting time at the telemedicine queue being 5.72 days shorter than the waiting time at the in-person queue. In comparison to this benchmark, employing an online triage tool to provide personalized predictions to patients can effectively reduce the average waiting time to 8.56 days (41% reduction). Furthermore, if the hospital implements a strict priority rule (with q^*) to incentivize patients to choose telemedicine, the average waiting time can be further reduced to 8.13 days (another 5% reduction). (Here, f satisfies the assumption of Case 1(a) in Proposition 7, so that the equilibrium is unique under with priority rule with $q^* = 1$.) In summary, both the provision of information alone and the combination of information with a priority rule demonstrate potential for improving the performance of the hospital's current practices.

8. Conclusion

This paper evaluates the potential advantages of investing in an online triage tool with underlying predictive analytics to facilitate telemedicine integration. Our approach involves developing a comprehensive framework that encompasses three key components: 1) a prediction model that forecasts patients' in-person follow-up needs after telemedicine service, 2) a queueing-game model that assesses the impact of information provision and prioritization on system performance, and 3)

Table 2 Equilibria Comparison ($p_{data} = 0.064$)

	Crude	Refined	First Best	Priority
	p^*	t^*	$\bar{t}$	q^*
Strategy	0.043	0.039	0.056	1

Table 3 Waiting Time Comparison (Unit: Days)

Average Waiting Time	Crude	Refined	First Best	Priority
Across All Patients	14.48	8.56	6.18	8.13
At The Telemedicine Queue	8.76	7.62	19.14	8.08
At The In-Person Queue	14.48	8.58	5.38	8.11

a case study that demonstrates the application of our theoretical results in a real-world setting. In terms of the value of information, our findings reveal that providing more information to patients has the advantage of maximizing the system stability region, but it does not consistently lead to a reduction in the average patient waiting time. Nevertheless, in the seemingly "problematic" situations where the refined information equilibrium yields a higher average waiting time than the crude information equilibrium, we show that by simultaneously implementing a priority rule, we can transform this information disadvantage into an advantage, achieving the optimal centralized system performance under specific conditions. To facilitate real-world implementation, we conduct a case study utilizing actual hospital data. The case study provides a step-wise demonstration of prediction model construction, model parameter calibration, and performance evaluation under the proposed policies.

We conclude by discussing several limitations of our work and identifying a few interesting avenues for future research.

First, we assume that patients have homogeneous and deterministic service value and waiting cost for analytical tractability. When heterogeneous and random service value and waiting cost are concerned, one promising direction is to extend our model by treating the service value and waiting cost as mappings of random patient types. By making additional assumptions regarding the linearity of these mappings, we conduct extensive numerical experiments and observe that the high-level structural insights, in terms of the impact of information provision and priority rules, remain robust. That said, it would be interesting to rigorously extend the model and analysis to incorporate random cost and reward structures that follow general distributions.

Second, our patients' decision model captures the core tradeoff between waiting time and treatment efficacy (which are the most important factors influencing patients' preferences reported by Mozes et al. (2022)) as they choose between visit modalities. That said, the real-world decision-making process for patients can be influenced by a multitude of other factors, including transportation time and costs, insurance coverage for different visit modalities, and potential repercussions of inadequate telemedicine treatment (such as clinical deterioration and mental burden). Incorporating additional elements into the decision model, such as introducing fixed costs for selecting an in-person visit or requiring in-person follow-up after telemedicine, is expected to be relatively straightforward and maintain the high-level insights. However, incorporating complex insurance policies and pricing mechanism is likely to necessitate fundamentally different development.

Third, in our research, we examine two ends of the information granularity spectrum: patients either possess complete or no information regarding their realized types, corresponding to the refined and crude information regimes, respectively. However, real-world prediction models, such as the logistic regression model developed in our case study, are susceptible to prediction errors. Moreover, patients may have partial understanding of their realized types and suitability for telemedicine. Therefore, a meaningful future research direction is to consider information granularity that falls between these two extremes on the spectrum.

References

Afeche P, Baron O, Milner J, Roet-Green R (2019) Pricing and prioritizing time-sensitive customers with heterogeneous demand rates. *Operations Research* 67(4):1184–1208.

Afeche P, Mendelson H (2004) Pricing and priority auctions in queueing systems with a generalized delay cost structure. *Management science* 50(7):869–882.

Akşin Z, Ata B, Emadi SM, Su CL (2017) Impact of delay announcements in call centers: An empirical approach. *Operations Research* 65(1):242–265.

Argon NT, Ziya S (2009) Priority assignment under imperfect information on customer type identities. *Manufacturing & Service Operations Management* 11(4):674–693.

Bavafa H, Hitt LM, Terwiesch C (2018) The impact of e-visits on visit frequencies and patient health: Evidence from primary care. *Management Science* 64(12):5461–5480.

Bavafa H, Savin S, Terwiesch C (2021) Customizing primary care delivery using e-visits. *Production and Operations Management* 30(11):4306–4327.

Çakıcı ÖE, Mills A (2022) Telehealth in acute care: Pay parity and patient access. *Available at SSRN 4045617* .

Crawford AM, Lightsey HM, Xiong GX, Striano BM, Greene N, Schoenfeld AJ, Simpson AK (2021) Interventional procedure plans generated by telemedicine visits in spine patients are rarely changed after in-person evaluation. *Regional Anesthesia & Pain Medicine* .

Cui S, Wang Z, Yang L (2023) A brief review of research on priority queues with self-interested customers. *Innovative Priority Mechanisms in Service Operations: Theory and Applications* 1–8.

Delana K, Deo S, Ramdas K, Subburaman GBB, Ravilla T (2023) Multichannel delivery in healthcare: the impact of telemedicine centers in southern india. *Management Science* 69(5):2568–2586.

Ding J, Freeman M, Hasija S (2022) Can predictive technology help improve acute care operations? investigating the impact of virtual triage adoption .

Edelson NM, Hilderbrand DK (1975) Congestion tolls for poisson queuing processes. *Econometrica: Journal of the Econometric Society* 81–92.

Eyrich NW, Andino JJ, Ukavwe RE, Farha MW, Patel AK, Triner D, Ellimoottil C (2022) The lack of a physical exam during new patient telehealth visits does not impact plans for office and operating room procedures. *Urology* 167:109–114.

Freeman M, Savva N, Scholtes S (2017) Gatekeepers at work: An empirical analysis of a maternity unit. *Management Science* 63(10):3147–3167.

Friedman AB SHBACABADGBJBR Gervasi S, As SM (2022) Telemedicine catches on: changes in the utilization of telemedicin e services during the covid-19 pandemic. *Am J Manag Care* 28(1).

Hassin R (1986) Consumer information in markets with random product quality: The case of queues and balking. *Econometrica: Journal of the Econometric Society* 1185–1195.

Hassin R (2016) *Rational queueing* (CRC press).

Hassin R, Haviv M (2003) *To queue or not to queue: Equilibrium behavior in queueing systems*, volume 59 (Springer Science & Business Media).

Hassin R, Roet-Green R (2017) The impact of inspection cost on equilibrium, revenue, and social welfare in a single-server queue. *Operations Research* 65(3):804–820.

Hassin R, Roet-Green R (2021) On queue-length information when customers travel to a queue. *Manufacturing & Service Operations Management* 23(4):989–1004.

Hatef E, Lans D, Bandeian S, Lasser EC, Goldsack J, Weiner JP (2022) Outcomes of in-person and telehealth ambulatory encounters during covid-19 within a large commercially insured cohort. *JAMA network open* 5(4):e228954–e228954.

Hathaway BA, Kagan E, Dada M (2023) The gatekeeper's dilemma:"when should i transfer this customer?". *Operations Research* 71(3):843–859.

Healthcare Dive (2022) Does telemedicine result in duplicative care? depends on the patient, study suggests. https://www.healthcaredive.com/news/does-telemedicine-result-in-duplicative-care-depends-on-the-patient-study/622773/.

Hu M, Momot R, Wang J (2022a) Privacy management in service systems. *Manufacturing & Service Operations Management* 24(5):2761–2779.

Hu Y, Chan CW, Dong J (2022b) Optimal scheduling of proactive service with customer deterioration and improvement. *Management Science* 68(4):2533–2578.

Hu Y, Chan CW, Dong J (2023) Prediction-driven surge planning with application in the emergency department. *Management Science* .

Ibrahim R (2018) Sharing delay information in service systems: a literature survey. *Queueing Systems* 89(1-2):49–79.

Kalwani NM, Johnson AN, Parameswaran V, Dash R, Rodriguez F (2021) Initial outcomes of cardioclick, a telehealth program for preventive cardiac care: observational study. *JMIR cardio* 5(2):e28246.

Kobeissi MM, Ruppert SD (2022) Remote patient triage: Shifting toward safer telehealth practice. *Journal of the American Association of Nurse Practitioners* 34(3):444.

Lekwijit TS, Song H, Terwiesch C, Chaiyachati K (2023) Multi-channel healthcare operations: The impact of video visits on the usage of in-person care. *Available at SSRN 4397550* .

Li M, Sun S, Gu W (2021) Tele-follow-up and outpatient care. *Available at SSRN 4499398* .

Liu N, Wang S, Zychlinski N (2023) Rl or url: Managing outpatient (tele) visits with strategic behavior. *Available at SSRN 4383199* .

Liu X, Goldenthal S, Li M, Nassiri S, Steppe E, Ellimoottil C (2021) Comparison of telemedicine versus in-person visits on impact of downstream utilization of care. *Telemedicine and e-Health* 27(10):1099–1104.

Mendelson H, Whang S (1990) Optimal incentive-compatible priority pricing for the m/m/1 queue. *Operations research* 38(5):870–883.

Meyer BC, Perrinez ES, Payne K, Carreño S, Partridge B, Braunlich B, Tangney J, Sylwestrzak M, Kremer B, Kane CJ, et al. (2023) Tele-untethered: Telemedicine without waiting rooms. *Quality Management in Health Care* 32(2):81.

Mihalj M, Carrel T, Gregoric ID, Andereggen L, Zinn PO, Doll D, Stueber F, Gabriel RA, Urman RD, Luedi MM (2020) Telemedicine for preoperative assessment during a covid-19 pandemic: Recommendations for clinical care. *Best Practice & Research Clinical Anaesthesiology* 34(2):345–351.

Mozes I, Mossinson D, Schilder H, Dvir D, Baron-Epel O, Heymann A (2022) Patients' preferences for telemedicine versus in-clinic consultation in primary care during the covid-19 pandemic. *BMC primary care* 23(1):33.

Naor P (1969) The regulation of queue size by levying tolls. *Econometrica: journal of the Econometric Society* 15–24.

Osmanlliu E, Kalwani NM, Parameswaran V, Qureshi L, Dash R, Scheinker D, Rodriguez F (2023) Sociodemographic disparities in the use of cardiovascular ambulatory care and telemedicine during the covid-19 pandemic. *American Heart Journal* .

Qin J, Chan CW, Dong J, Homma S, Ye S (2023a) Telemedicine is associated with reduced socioeconomic disparities in outpatient clinic no-show rates. *Journal of Telemedicine and Telecare* 1357633X231154945.

Qin J, Chan CW, Dong J, Homma S, Ye S (2023b) Waiting online versus in-person: An empirical study on outpatient clinic visit incompletion .

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.

Roet-Green R, Shetty A (2022) On designing a socially optimal expedited service and its impact on individual welfare. *Manufacturing & Service Operations Management* 24(3):1843–1858.

Saghafian S, Hopp WJ, Iravani SM, Cheng Y, Diermeier D (2018) Workload management in telemedical physician triage and other knowledge-based service systems. *Management Science* 64(11):5180–5197.

Shah VV, Villaflores CW, Chuong LH, Leuchter RK, Kilaru AS, Vangala S, Sarkisian CA (2022) Association between in-person vs telehealth follow-up and rates of repeated hospital visits among patients seen in the emergency department. *JAMA Network Open* 5(10):e2237783–e2237783.

Shumsky RA, Pinker EJ (2003) Gatekeepers and referrals in services. *Management Science* 49(7):839–856.

Singh S, Gurvich I, Van Mieghem J (2024) Feature-driven priority queuing .

Srinivasan M, Asch S, Vilendrer S, Thomas SC, Bajra R, Barman L, Edwards LM, Filipowicz H, Giang L, Jee O, et al. (2020) Qualitative assessment of rapid system transformation to primary care video visits at an academic medical center. *Annals of internal medicine* 173(7):527–535.

Staats BR, Dai H, Hofmann D, Milkman KL (2017) Motivating process compliance through individual electronic monitoring: An empirical examination of hand hygiene in healthcare. *Management Science* 63(5):1563–1585.

SteelFisher GK, McMurtry CL, Caporello H, Lubell KM, Koonin LM, Neri AJ, Ben-Porath EN, Mehrotra A, McGowan E, Espino LC, et al. (2023) Video telemedicine experiences in covid-19 were positive, but physicians and patients prefer in-person care for the future: Study examines patient and physician opinion of telemedicine experiences during covid-19. *Health Affairs* 42(4):575–584.

Sun S, Wang G (2021) Does telehealth reduce rural-urban care access disparities? evidence from covid-19 telehealth expansion. *Evidence from COVID-19 Telehealth Expansion (February 18, 2021)* .

Sun Z, Argon NT, Ziya S (2022) When to triage in service systems with hidden customer class identities? *Production and Operations Management* 31(1):172–193.

Sunar N, Staats BR (2022) Telemedicine for inclusive care: Remedy for socioeconomic health disparities? *Available at SSRN 4103887* .

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, Zhang Y, Zhou YP (2022) Delay information in virtual queues: A large-scale field experiment on a major ride-sharing platform. *Management Science* 68(8):5745–5757.

Appendix A: Case Study Supplements

Table A.1 below provides the summary statistics of the features used in the logistic regression model for the focal data set.

Table A.1 Summary Statistics of the Data Set

Characteristic	Statistics by Care Modality, No. (%)		
	Total (n = 3275)	In-person (n = 3065)	Telemedicine (n = 210)
Unique Patients	3159	2969	209
Days to Follow-up, Mean (SD)	50.11 (145.57)	53.11 (150.00)	6.35 (5.44)
Patient Age Group			
Under 19	279 (9)	279 (9)	0 (0)
$[20, 39]$	480 (15)	464 (15)	16 (8)
$[40, 59]$	1224 (37)	1127 (37)	97 (46)
$[60, 79]$	1146 (35)	1065 (35)	81 (39)
Above 80	146 (4)	130 (4)	16 (8)
Sex			
Male	1206 (37)	1203 (39)	3 (1)
Female	2069 (63)	1862 (61)	207 (99)
Ethnicity			
Not Hispanic or Latinx	3061 (93)	2859 (93)	202 (96)
Hispanic or Latinx	174 (5)	166 (5)	8 (4)
Unknown	40 (1)	40 (1)	0 (0)
Race			
White	1952 (60)	1829 (60)	123 (59)
Black	769 (23)	727 (24)	42 (20)
Asian	288 (9)	254 (8)	34 (16)
Native American	3 (0)	3 (0)	0 (0)
Hawaiian	3 (0)	3 (0)	0 (0)
Unknown	260 (8)	249 (8)	11 (5)
Primary Insurance			
Commercial	2035 (62)	1899 (62)	136 (65)
Medicare	742 (23)	686 (22)	56 (27)
Medicaid	231 (7)	229 (7)	2 (1)
Self Pay	64 (2)	64 (2)	0 (0)
Other Insurance	203 (6)	187 (6)	16 (8)
Source of Patients			
Home	2093 (64)	1883 (61)	210 (100)
Clinic or Physician Office	1182 (36)	1182 (39)	0 (0)
Number of Visits in 2020, Mean (SD)	0.34 (0.98)	0.35 (1.00)	0.19 (0.54)
Charlson Comorbidity Index, Mean (SD)	0.14 (0.53)	0.07 (0.38)	1.11 (1.11)

Table A.2 and A.3 provide the rest of the logistic regression results that are not shown in the main body of the paper. Table A.2 shows the effect of diagnosis code and Table A.3 shows the impact of race, and the year and season effect.

Table A.2 Regression Results (Part 2)

	followup7	followup14	followup21	followup30
	(1)	(2)	(3)	(4)
A00-B99	-25.617	-25.542	-25.469	-22.443
	(131010.700)	(131010.690)	(131010.683)	(29232.438)
C00-D49	0.339	1.311^{***}	1.053^{**}	1.047^{**}
	(0.328)	(0.409)	(0.414)	(0.416)
D50-D89	-23.901	-24.000	-24.242	-21.464
	(59267.355)	(60097.147)	(58872.173)	(12974.709)
F01-F99	1.128	0.997	0.906	0.809
	(1.788)	(1.677)	(1.672)	(1.639)
G00-G99	-1.328	-1.438	-1.350	-1.528
	(1.122)	(1.111)	(1.104)	(1.109)
H60-H95	0.050	-0.689	-1.038	-1.054
	(1.665)	(2.101)	(2.405)	(2.394)
I00-I99	-0.101	0.052	-0.178	0.049
	(0.474)	(0.447)	(0.454)	(0.443)
J00-J99	-0.793	-0.743	-0.906	-0.537
	(0.728)	(0.721)	(0.729)	(0.666)
L00-L99	0.097	-0.168	-0.236	0.528
	(0.679)	(0.660)	(0.657)	(0.613)
M00-M99	-0.006	0.282	0.144	0.193
	(0.353)	(0.331)	(0.332)	(0.324)
N00-N99	1.123^{***}	1.101^{***}	1.062^{***}	1.243^{***}
	(0.357)	(0.377)	(0.377)	(0.385)
O00-09A	0.065	-0.215	-0.422	-0.588
	(1.386)	(1.471)	(1.561)	(1.663)
P00-P96	0.000	-0.000	-0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Q00-Q99	-1.519^{*}	-0.913	-1.000	-1.217^{*}
	(0.831)	(0.687)	(0.680)	(0.699)
R00-R99	-2.213^{***}	-1.325^{***}	-1.193^{***}	-1.140^{***}
	(0.453)	(0.373)	(0.362)	(0.352)
S00-T88	-24.597	-1.217	-1.236	-1.843
	(59093.094)	(1.238)	(1.232)	(1.223)
U00-U85	-2.969^{***}	-0.197	-0.491	-0.269
	(1.086)	(0.633)	(0.645)	(0.647)
V00-Y99	24.930	1.685	1.604	2.046
	(59093.094)	(1.409)	(1.407)	(1.410)
Z00-Z99	-1.034^{***}	-0.650^{***}	-0.701^{***}	-0.768^{***}
	(0.198)	(0.213)	(0.222)	(0.225)
Observations	3275	3275	3275	3275

Note: *p<0.1; **p<0.05; ***p<0.01

Table A.3 Regression Results (Part 3)

	followup7 (1)	followup14 (2)	followup21 (3)	followup30 (4)
Race (Asian)	-0.255	-0.276*	-0.265	-0.296*
	(0.157)	(0.162)	(0.163)	(0.163)
Race (Black)	0.001	-0.048	0.010	0.017
	(0.103)	(0.105)	(0.104)	(0.103)
Race (Havaiian)	1.533	1.270	1.177	1.124
	(1.257)	(1.237)	(1.236)	(1.236)
Race (Native American)	1.542	1.524	1.458	1.378
	(1.278)	(1.273)	(1.268)	(1.266)
Race (Unknown)	-0.001	0.006	-0.014	0.045
	(0.191)	(0.193)	(0.193)	(0.190)
Year&Quarter (2021q2)	0.069	0.057	0.026	0.002
	(0.113)	(0.112)	(0.111)	(0.109)
Year&Quarter (2021q3)	0.836***	0.749***	0.676***	0.618***
	(0.134)	(0.135)	(0.134)	(0.133)
Year&Quarter (2021q4)	1.710***	1.991***	2.097***	2.140***
	(0.279)	(0.312)	(0.329)	(0.340)
Year&Quarter (2022q1)	1.049***	0.926***	1.218***	1.283***
	(0.325)	(0.358)	(0.382)	(0.391)
Year&Quarter (2022q2)	0.657	0.914*	1.237**	1.448**
	(0.422)	(0.511)	(0.573)	(0.613)
Year&Quarter (2022q3)	1.030**	1.891***	2.555***	2.457***
	(0.443)	(0.591)	(0.777)	(0.777)
Year&Quarter (2022q4)	1.490***	2.705***	3.128***	3.019***
	(0.370)	(0.520)	(0.639)	(0.640)
Year&Quarter (2023q1)	1.882***	2.403***	2.965***	3.126***
	(0.408)	(0.467)	(0.578)	(0.643)
Year&Quarter (2023q2)	0.834***	1.694***	1.791***	1.633***
	(0.310)	(0.398)	(0.428)	(0.430)
Year&Quarter (2023q3)	0.694*	1.102**	0.937**	0.803*
	(0.412)	(0.467)	(0.476)	(0.478)
Observations	3275	3275	3275	3275

Note: *p<0.1; **p<0.05; ***p<0.01

Appendix B: Proofs of Main Results

B.1. Proof of Proposition 1

Proof: We first derive the stability condition for the crude patient strategy. To ensure the crude system is stable, by Lemma 1, the following two equations have to be satisfied for some $p \in [0,1]$:

$$\mu_1 - \lambda p > 0 \quad \Leftrightarrow \quad p < \frac{\mu_1}{\lambda},$$

$$\mu_2 - \lambda(1 - p + p\mathbb{E}[X]) > 0 \quad \Leftrightarrow \quad \lambda p(1 - \mathbb{E}[X]) > \lambda - \mu_2 \quad \Leftrightarrow \quad p > \frac{\lambda - \mu_2}{\lambda(1 - \mathbb{E}[X])} := p_s.$$

We divide our discussion into 4 cases.

Case 1. When $\mu_1 > \lambda$ and $\mu_2 > \lambda$, all $p \in [0,1]$ satisfy the above two inequalities.

Case 2. When $\mu_1 > \lambda$ and $\lambda\mathbb{E}[X] < \mu_2 \leq \lambda$, all $p \in (p_s, 1]$ satisfy the above two inequalities.

Case 3. When $\mu_1 \leq \lambda$ and $\mu_2 > \lambda$, all $t \in [0, \mu_1/\lambda)$ satisfy the above two inequalities.

Case 4. When $\mu_1 \leq \lambda$ and $\lambda\mathbb{E}[X] < \mu_2 \leq \lambda$, if $p_s = (\lambda - \mu_2)/(\lambda(1 - \mathbb{E}[X])) < \mu_1/\lambda \Leftrightarrow \lambda - (1 - \mathbb{E}[X])\mu_1 < \mu_2$, we have all $p \in (p_s, \mu_1/\lambda)$ satisfying the above two inequalities. Otherwise, there is no $p \in [0,1]$ satisfying the above two inequalities.

To summarize, the stability condition for the crude system is $\mu_2 > \lambda\mathbb{E}[X]$ when $\mu_1 > \lambda$, and $\mu_2 > \lambda - (1 - \mathbb{E}[X])\mu_1$ when $\mu_1 \leq \lambda$. The set of p satisfying the stability condition can be summarized as the interval $(\max(0, p_s), \min(1, \mu_1/\lambda))$.

Given Equation (5), any strictly-mixed strategy $g(U) = p \in (0,1)$ should solve the following equation:

$$\frac{1}{\mu_1 - \lambda p} + \frac{\mathbb{E}[X]}{\mu_2 - \lambda(1 - p + p\mathbb{E}[X])} = \frac{1}{\mu_2 - \lambda(1 - p + p\mathbb{E}[X])}. \tag{B.1}$$

The left-hand side (LHS) of Equation (B.1) is the expected waiting time at the telemedicine queue and the right-hand side (RHS) of Equation (B.1) is the expected waiting time at the in-person queue. To better compare the two sides, we rewrite Equation (B.1) as

$$\frac{1}{\mu_1 - \lambda p} = \frac{1 - \mathbb{E}[X]}{\mu_2 - \lambda(1 - p + p\mathbb{E}[X])}. \tag{B.2}$$

By taking the derivative it is straightforward to show that the LHS of Equation (B.2) is strictly increasing with respect to p, and the RHS of Equation (B.2) is strictly decreasing with respect to p. This implies that there is at most one root (if any) of p^* that solves Equation (B.2), and thus Equation (B.1). We next characterize patients' equilibrium strategies using Equation (B.2). We divide the discussion into the following three cases which partition the parameter space. More specifically, a pure strategy equilibrium exists in Cases 1 and 2, and a mixed strategy equilibrium exists in Case 3.

Case 1. When $\lambda + (1 - \mathbb{E}[X])\mu_1 = \mu_2$, we have LHS$= 1/\mu_1 = (1 - \mathbb{E}[X])/(\mu_2 - \lambda) =$ RHS at $p = 0$, thus $p^* = 0$ is the only solution to Equations (B.2) and (B.1), which implies $p^* = 0$ is the equilibrium. When $\lambda + (1 - \mathbb{E}[X])\mu_1 < \mu_2$, LHS$= 1/\mu_1 > (1 - \mathbb{E}[X])/(\mu_2 - \lambda) =$ RHS $\Rightarrow$ LHS>RHS for all p in Equation (B.2). This implies that in Equation (B.1), we also have LHS>RHS for all p, because the comparison direction does not change when we subtract the same number from both sides. It follows that $p^* = 0$ is the equilibrium. To summarize both scenarios, when $\lambda + (1 - \mathbb{E}[X])\mu_1 \leq \mu_2$, $p^* = 0$ is the equilibrium.

Case 2. If $\mu_1 > \lambda$, when $\mu_2 = (1 - \mathbb{E}[X])\mu_1 - (1 - 2\mathbb{E}[X])\lambda$, we have LHS$= 1/(\mu_1 - \lambda) =$

$(1-\mathbb{E}[X])/(\mu_2-\lambda\mathbb{E}[X])$ =RHS at $p=1$, thus $p^*=1$ is the only solution to Equations (B.2) and (B.1), which implies that $p^*=1$ is the equilibrium. When $\mu_2<(1-\mathbb{E}[X])\mu_1-(1-2\mathbb{E}[X])\lambda$, LHS$=1/(\mu_1-\lambda)<(1-\mathbb{E}[X])/(\mu_2-\lambda\mathbb{E}[X])$ =RHS $\Rightarrow$ LHS<RHS for all p in Equation (B.2). Hence, in Equation (B.1), we also have LHS<RHS for all p, because the comparison direction does not change if we subtract the same number from both sides. It follows that $p^*=1$ is the equilibrium. In conclusion, when $\mu_2\leq(1-\mathbb{E}[X])\mu_1-(1-2\mathbb{E}[X])\lambda$ and $\mu_1>\lambda$, $p^*=1$ is the equilibrium.

Case 3. Otherwise, the LHS and RHS of Equation (B.2) must intersect and there is a unique solution $p^*\in(0,1)$. Solving Equation (B.2) is equivalent to solving

$$(1-\mathbb{E}[X])(\mu_1-\lambda p)=\mu_2-\lambda(1-p+p\mathbb{E}[X]). \tag{B.3}$$

Solving the above equation, we get a unique solution $p^*=((1-\mathbb{E}[X])\mu_1+\lambda-\mu_2)/(2\lambda(1-\mathbb{E}[X]))$ which is also the only solution to Equation (B.1). To check whether p^* satisfies the crude stability condition, we plug p^* into $\mu_1-\lambda p^*>0$, which gives $\mu_2>\lambda-(1-\mathbb{E}[X])\mu_1$, the crude stability condition. Q.E.D.

B.2. Proof of Lemma 2

Proof: Suppose the equilibrium is of mixed strategy equilibrium for every type of patient. Then by definition, the expected waiting times at both queues should be the same for all types. That is, let $g:[0,1]\to[0,1]$ denote any patient strategy and we denote $p=\int_0^1 g(x)dx$, then we should have

$$\frac{1}{\mu_1-\lambda p}+\frac{f(t)}{\mu_2-\lambda(1-p+D(p))}=\frac{1}{\mu_2-\lambda(1-p+D(p))},\quad \forall t\in[0,1], \tag{B.4}$$

where $D(p)=\int_0^1 g(x)f(x)dx$. The LHS of Equation (B.4) is the expected waiting time at the telemedicine queue, and the RHS of Equation (B.4) is the expected waiting time at the in-person queue. It is impossible that the above equation holds for all $t\in[0,1]$ because the denominators are independent of t and the value of $f(t)$ is unique for each $t\in[0,1]$. Rearranging the terms of Equation (B.4), we get

$$\frac{1}{\mu_1-\lambda p}=\frac{1-f(t)}{\mu_2-\lambda(1-p+D(p))}. \tag{B.5}$$

For a fixed strategy g, the LHS of Equation (B.5) is a constant while the RHS of Equation (B.5) is a strictly decreasing function, thus if the two lines intersect, there is only one intersection, which we denote as t^*. We consider the following cases for that given g.

Case 1. If $t^*\in[0,1]$ exists under strategy g, then because f is strictly increasing, in Equation (B.5), we have LHS<RHS for all $t<t^*$, and LHS>RHS for all $t>t^*$. Thus in Equation (B.4), we also have LHS<RHS for all $t<t^*$, and LHS>RHS for all $t>t^*$ because adding the same value on both sides does not change the order of the inequality. That is, all types $t<t^*$ deviate to the telemedicine service, and all types $t<t^*$ deviate to the in-person service. Thus $g(t)=1$ for $t\leq t^*$, $g(t)=0$ for $t>t^*$, and $p=t^*$ is the equilibrium.

Case 2. If in Equation (B.5), LHS<RHS for all $t\in[0,1]$ under strategy g, which implies in Equation (B.4), LHS<RHS for all $t\in[0,1]$ under strategy g. Then we define $t^*=1$, and $g(t)=1$ for $t\leq t^*$ is the equilibrium.

Case 3. If in Equation (B.5), LHS>RHS for all $t\in[0,1]$ under strategy g, which implies in Equation (B.4), LHS>RHS for all $t\in[0,1]$ under strategy g. Then we define $t^*=0$, and $g(t)=0$ for $t\geq t^*$ is the equilibrium.

Q.E.D.

B.3. Proof of Proposition 2

Proof: We prove that in the refined system, a unique equilibrium threshold t^* exists. We first derive the refined stability condition for the one-threshold type strategy. Notice that under the one-threshold type of strategy, p reduces to t and $D(p)$ reduces to $F(t)$ if t is the threshold. Then to ensure the refined system is stable, by Lemma 1, the following two equations have to be satisfied for some $t \in [0,1]$:

$$\mu_1 - \lambda t > 0 \Leftrightarrow t < \frac{\mu_1}{\lambda},$$

$$\mu_2 - \lambda(1 - t + F(t)) > 0.$$

We divide our discussion into 4 cases.

Case 1. When $\mu_1 > \lambda$ and $\mu_2 > \lambda$, because $1 - t + F(t)$ is strictly decreasing in t with the upper bound equal to 1 since $F(0) = 0$, all $t \in [0,1]$ satisfy the above two inequalities.

Case 2. When $\mu_1 > \lambda$ and $\lambda\mathbb{E}[X] < \mu_2 \le \lambda$, we define $t_s := \min\{t \in [0,1) | \mu_2 - \lambda(1-t+F(t)) = 0 \text{ and } \lambda\mathbb{E}[X] < \mu_2 \le \lambda\}$. Then because $1 - t + F(t)$ is strictly decreasing in t with the lower bound $1 - 1 + F(1) = \mathbb{E}[X]$, all $t \in (t_s, 1]$ satisfy the above two inequalities.

Case 3. When $\mu_1 \le \lambda$ and $\mu_2 > \lambda$, all $t \in [0, \mu_1/\lambda)$ satisfy the above two inequalities.

Case 4. When $\mu_1 \le \lambda$ and $\lambda\mathbb{E}[X] < \mu_2 \le \lambda$, if $\mu_2 - \lambda(1 - \mu_1/\lambda + F(\mu_1/\lambda)) > 0$, we have $t_s < \mu_1/\lambda$ and all $t \in (t_s, \mu_1/\lambda)$ satisfying the above two inequalities. Otherwise, there is no $t \in [0,1]$ satisfying the above two inequalities.

To summarize, the stability condition for the refined system is $\mu_2 > \lambda\mathbb{E}[X]$ when $\mu_1 > \lambda$, and $\mu_2 > \lambda(1 - \mu_1/\lambda + F(\mu_1/\lambda))$ when $\mu_1 \le \lambda$. The set of t satisfying the stability condition can be summarized as the interval $(\max(0, t_s), \min(1, \mu_1/\lambda))$.

Continuing from the Proof of Lemme 2, we then prove that under the refined stability condition, the threshold t^* always exists and it is unique. Notice that for all strategy g under which the intersection t^* exists, t^* has to solve

$$\frac{1}{\mu_1 - \lambda t} = \frac{1 - f(t)}{\mu_2 - \lambda(1 - t + F(t))}. \tag{B.6}$$

We show that Equation (B.6) has at most one solution of t. The derivative of the RHS of Equation (B.6) with respect to t is

$$\frac{-f'(t)(\mu_2 - \lambda(1 - t + F(t))) - (1 - f(t))^2\lambda}{(\mu_2 - \lambda(1 - t + F(t)))^2} < 0$$

because $f'(t) > 0$. Thus we have the RHS strictly decreasing with respect to t. Obviously, the LHS strictly increasing with respect to t. So if the two lines intersect, they intersect at most once, which implies that the solution t^* is unique if it exists and thus the corresponding equilibrium strategy g is unique.

We characterize the value of t^* in different situations.

Case 1. $t^* = 1$ is never an equilibrium. Because to have $t^* = 1$, by (B.6), we need to ensure LHS≤RHS$= 0$ at $t = 1$, which is equivalent to $\mu_1 - \lambda t^* \le 0$, this violates the stability condition.

Case 2. If $\lambda + \mu_1 \le \mu_2$, $t^* = 0$ is the equilibrium. This is because $t^* = 0$ if LHS≥RHS at $t = 0$, which is equivalent to $\mu_2 - (1 - F(0))\lambda \ge (1 - f(0))\mu_1 \Leftrightarrow \lambda + \mu_1 \le \mu_2$, because $F(0) = f(0) = 0$.

Case 3. Otherwise, a unique solution $t^* \in (0,1)$ to Equation (B.6) exists because the LHS and RHS must intersect and they intersect only once by the monotonicity. To show the existence of the intersection, we first show that at the maximum $t = \min(1, \mu_1/\lambda)$, we have LHS>RHS. The LHS equals $1/\mu_1 - \lambda > 0$ when $t = 1$, and the RHS equals 0 when $t = 1$. The LHS→ ∞ when $t \to \mu_1/\lambda$, and the RHS< $1/(\mu_2 - \lambda) < \infty$ when $t \to \mu_1/\lambda$. We then show at the minimum of $t = \max(0, t_s)$, we have LHS<RHS. Because the condition in Case 3 is already the complement of condition in Case 2, we have LHS<RHS at $t = 0$ if $t = 0$ is feasible. At $t = t_s$, the $RHS \to \infty$, and the $LHS < \infty$ because $LHS \to \infty$ only when $t \to \mu_1/\lambda$. As long as $t_s < \mu_1/\lambda$, which is ensured by the stability condition, we have $LHS < \infty$ for each specific t_s. Q.E.D.

B.4. Sensitivity of the Equilibria

Lemma B.1 *Let p^* be the probability in Proposition 1 that characterizes the unique patient equilibrium in the crude information regime, and let t^* be the threshold in Proposition 2 that characterizes the unique patient equilibrium in the refined information regime. The following results hold:*

1. *For $(\lambda, \mu_1, \mu_2, f) \in \mathcal{S}^{crude}$, the probability p^* increases with respect to μ_1 and decreases with respect to μ_2;*
2. *For $(\lambda, \mu_1, \mu_2, f) \in \mathcal{S}^{refined}$, the threshold t^* increases with respect to μ_1 and decreases with respect to μ_2.*

Proof of Lemma B.1: We start by analyzing the crude equilibrium. According to Proposition 1, when $\lambda + (1 - \mathbb{E}[X])\mu_1 \le \mu_2$, $p^* = 0$; when $\mu_2 \le (1 - \mathbb{E}[X])\mu_1 - (1 - 2\mathbb{E}[X])\lambda$ and $\mu_1 > \lambda$, $p^* = 1$; otherwise, $p^* = ((1 - \mathbb{E}[X])\mu_1 + \lambda - \mu_2)/(2\lambda(1 - \mathbb{E}[X]))$. Obviously, $p^* = ((1 - \mathbb{E}[X])\mu_1 + \lambda - \mu_2)/(2\lambda(1 - \mathbb{E}[X]))$ increases with respect to μ_1 and decreases with respect to μ_2. In addition, when $\lambda + (1 - \mathbb{E}[X])\mu_1 \le \mu_2$, increasing μ_1 or decreasing μ_2 to some extent violates this condition and thus forces p^* to become positive. When $\mu_2 \le (1 - \mathbb{E}[X])\mu_1 - (1 - 2\mathbb{E}[X])\lambda$, increasing μ_1 or decreasing μ_2 maintains this condition and keeps $p^* = 1$.

For the refined equilibrium, according to Proposition 2, when $\lambda + \mu_1 \le \mu_2$, $t^* = 0$. Increasing μ_1 or decreasing μ_2 to some extent violates the condition of $\lambda + \mu_1 \le \mu_2$ and thus forces t^* to become positive.

Otherwise, for $0 < t^* < 1$, rearranging (B.7) and plug in $t = t^*$, we get

$$\mu_2 - \lambda(1 - t^* + F(t^*)) - (1 - f(t^*))(\mu_1 - \lambda t^*) = 0. \tag{B.7}$$

Take the derivative of (B.7) with respect to μ_1 on both sides, we get

$$-\lambda\left(-\frac{\partial t^*}{\partial \mu_1} + f(t^*)\frac{\partial t^*}{\partial \mu_1}\right) + (\mu_1 - \lambda t^*)f'(t^*)\frac{\partial t^*}{\partial \mu_1} - (1 - f(t^*))\left(1 - \lambda\frac{\partial t^*}{\partial \mu_1}\right) = 0.$$

Simplify it and we get

$$((\mu_1 - \lambda t^*)f'(t^*) + 2\lambda(1 - f(t^*)))\frac{\partial t^*}{\partial \mu_1} = 1 - f(t^*).$$

Because $(\mu_1 - \lambda t^*)f'(t^*) + 2\lambda(1 - f(t^*)) > 0$ and $1 - f(t^*) > 0$, we have $\frac{\partial t^*}{\partial \mu_1} > 0$.

Take the derivative of (B.7) with respect to μ_2 on both sides, we get

$$1 - \lambda\left(-\frac{\partial t^*}{\partial \mu_2} + f(t^*)\frac{\partial t^*}{\partial \mu_2}\right) + (\mu_1 - \lambda t^*)f'(t^*)\frac{\partial t^*}{\partial \mu_2} + \lambda(1 - f(t^*))\frac{\partial t^*}{\partial \mu_2} = 0.$$

Simplify it and we get

$$((\mu_1 - \lambda t^*)f'(t^*) + 2\lambda(1 - f(t^*)))\frac{\partial t^*}{\partial \mu_2} = -1. \tag{B.8}$$

Because $(\mu_1 - \lambda t^*)f'(t^*) + 2\lambda(1 - f(t^*)) > 0$, we have $\frac{\partial t^*}{\partial \mu_2} < 0$. Q.E.D.

B.5. Proof of Proposition 3

Proof: We show that for fixed f, if the stability condition for the refined system, $\mu_2 > \lambda - \mu_1 + \lambda F(\mu_1/\lambda)$ if $\mu_1 \le \lambda$, holds, then the stability condition for the crude system, $\mu_2 > \lambda - \mu_1 + \mu_1 \mathbb{E}[X]$, also holds. We consider two cases:

Case 1. When $\mu_1 > \lambda$, the refine system is always stable. For the crude system, we have $\lambda - \mu_1 + \mu_1\mathbb{E}[X] < \lambda\mathbb{E}[X] < \mu_2$, where the second inequality follows from Assumption (1). Therefore, the conditions characterizing the two sets above both reduce to $\mu_2 > \lambda\mathbb{E}[X]$, which always holds by Assumption (1).

Case 2. When $\mu_1 \le \lambda$, it is sufficient to show that $\mu_1\mathbb{E}[X] \ge \lambda F(\mu_1/\lambda)$, which is equivalent to $(\mu_1/\lambda)\mathbb{E}[X] - F(\mu_1/\lambda) \ge 0$. Let $p = \mu_1/\lambda$, then $0 < p \le 1$. Then because $p\mathbb{E}[X] = p\int_0^1 f(x)dx$ and $F(p) = \int_0^p f(x)dx = p\int_0^1 f(px)dx$, it follows that $p\mathbb{E}[X] - F(p) = p\int_0^1 (f(x) - f(px))dx > 0$ since f is strictly increasing. Q.E.D.

B.6. Proof of Theorem 1 and Proposition 4

Proof: The proof is presented by firstly fixing λ, f, and μ_1, and varying the value of μ_2. After deriving the expression that calculates $\bar{\mu}_2$ and $\tilde{\mu}_2$, we fix f and λ, and show that $\bar{\mu}_2$ and $\tilde{\mu}_2$ are continuous with respect to μ_1. Fixing λ, f, and μ_1:

Case 1. According to Proposition 1 and Proposition 2, when $\mu_2 \ge \mu_1 + \lambda$, $p^* = t^* = 0$, so $h^c(p^*) = h^r(t^*)$.

Case 2a. From now on, within each case, we further divide our discussion into subcases where in the first subcase, we conduct the equilibrium comparison, and in the second subcase, we conduct the average waiting time comparison. When $\mu_2 < \mu_1 + \lambda$,

2a.1 (Equilibrium Comparison) Our first goal is to show that there exists a $\bar{\mu}_2$ such that for $\mu_2 \ge \bar{\mu}_2$, we have $p^* \le t^*$. When $\mu_2 \ge (1 - \mathbb{E}[X])\mu_1 + \lambda$, by Proposition 1 and Proposition 2, $0 = p^* < t^*$. When $\mu_2 < (1 - \mathbb{E}[X])\mu_1 + \lambda$, Lemma B.2 below shows that at the μ_2^* under which $t^* = f^{-1}(\mathbb{E}[X])$, we have $p^* > t^*$. Because by the proof of Lemma B.1, when $p^* \ne 0$, both t^* and p^* are continuous and strictly increasing as μ_2 decreases, there must exist a $\bar{\mu}_2$ such that at $\mu_2 = \bar{\mu}_2$ $p^* = t^*$, and for $\mu_2 \in [\bar{\mu}_2, (1 - \mathbb{E}[X])\mu_1 + \lambda)$, $0 < p^* \le t^*$.

Lemma B.2 *At* $\mu_2 = \mu_2^* := \lambda(1 - t^* + F(t^*)) + (1 - f(t^*))(\mu_1 - \lambda t^*)$ *where* $t^* = f^{-1}(\mathbb{E}[X])$, $p^* > t^*$.

The proof of Lemma (B.2) is provided in Section B.6.1.

Notice that $\bar{\mu}_2$ can be obtained by solving equation $(\mu_1 - \lambda p^*)(1 - f(p^*)) = \mu_2 - \lambda(1 - p^* + F(p^*))$, where $p^* = ((1 - \mathbb{E}[X])\mu_1 + \lambda - \mu_2)/(2\lambda(1 - \mathbb{E}[X]))$. If we fix λ and f, then $\bar{\mu}_2$ is continuous with respect to μ_1 because both equations are continuous in μ_1. So we denote it as $\bar{\mu}_2(\mu_1)$.

2a.2 (Average Waiting Time Comparison) We then compare the average waiting time. We show when $\bar{\mu}_2 \le \mu_2 < \mu_1 + \lambda$, $h^c(p^*) > h^r(t^*)$. By the definition of the average waiting time, p^*, and t^*, we have

$$h^c(p^*) = \frac{p^*}{\mu_1 - \lambda p^*} + \frac{1 - p^* + p^*\mathbb{E}[X]}{\mu_2 - \lambda(1 - p^* + p^*\mathbb{E}[X])} = \frac{1}{\mu_2 - \lambda(1 - p^* + p^*\mathbb{E}[X])},$$

where the second equality is because (B.3), and

$$h^r(t^*) = \frac{t^*}{\mu_1 - \lambda t^*} + \frac{1 - t^* + F(t^*)}{\mu_2 - \lambda(1 - t^* + F(t^*))} = \frac{1 + F(t^*) - t^* f(t^*)}{\mu_2 - \lambda(1 - t^* + F(t^*))},$$

where the second equality is because (B.7).

When $\bar{\mu}_2 \le \mu_2 < \mu_1 + \lambda$, Case 2a.1 has shown $p^* \le t^*$, then, we have $1 - p^* + p^*\mathbb{E}[X] \ge 1 - p^* + F(p^*) \ge 1 - t^* + F(t^*)$ because $1 - t + F(t)$ decreases in t. Thus the order of the denominators is $\mu_2 - \lambda(1 - p^* + p^*\mathbb{E}[X]) \le \mu_2 - \lambda(1 - t^* + F(t^*))$. In addition, because term $1 + F(t) - tf(t)$ also decreases with respect to t, which can be proved by taking the derivative, the order of the numerators is $1 > 1 + F(t^*) - t^* f(t^*)$ because $t^* > 0$. So that $h^c(p^*) > h^r(t^*)$.

Case 2b and Case 3.

2b.1 and 3.1 (Equilibrium Comparison) To show that when $\mu_2 < \bar{\mu}_2$, $p^* > t^*$, we consider two sub-cases: $\mu_2^* < \mu_2 < \bar{\mu}_2$ and $\mu_2 < \mu_2^*$. For both cases, we show that $p^* > t^*$.

(a) When $\mu_2^* < \mu_2 < \bar{\mu}_2$, because t^* is strictly decreasing in μ_2, we have $t^* < f^{-1}(\mathbb{E}[X])$. We show that $t^* - p^*$ is monotonically increasing with respect to μ_2. By the proof of Lemma B.1,

$$\frac{\partial(t^* - p^*)}{\partial \mu_2} = \frac{\partial t^*}{\partial \mu_2} - \frac{\partial p^*}{\partial \mu_2} = \frac{-1}{2\lambda(1 - f(t^*)) + (\mu_1 - \lambda t^*)f'(t^*)} + \frac{1}{2\lambda(1 - \mathbb{E}[X])}.$$

Notice the first term of the derivative is negative and the second term is positive. But in the current case, $t^* < f^{-1}(\mathbb{E}[X]) \Rightarrow f(t^*) < \mathbb{E}[X] \Rightarrow 2\lambda(1 - f(t^*)) > 2\lambda(1 - \mathbb{E}[X]) \Rightarrow 1/(2\lambda(1 - f(t^*)) + (\mu_1 - \lambda t^*)f'(t^*)) < 1/(2\lambda(1 - \mathbb{E}[X]))$. Therefore, the derivative is positive.

(b) When $\mu_2 < \mu_2^*$, i.e. $t^* > f^{-1}(\mathbb{E}[X])$, Lemma B.3 below show that when $p^* \neq 1$, $p^* > t^*$. When $p^* = 1$, by Proposition 2, $t^* < 1$ always holds, thus we still have $t^* < p^*$.

Lemma B.3 *When $t^* > f^{-1}(\mathbb{E}[X])$ and $p^* \neq 1$, $p^* > t^*$.*

The proof of Lemma (B.3) is provided in Section B.6.1. To summarize, the result follows from Part (a), Part (b), and Lemma B.2.

2b.2 and 3.2 (Average Waiting Time Comparison) When $\mu_2 < \bar{\mu}_2$, we let $\mathcal{K}$ represents a set that contains all μ_2 at which $h^c(p^*) = h^r(t^*)$. Then, every element of set $\mathcal{K}$ (if $\mathcal{K}$ is not empty) is a finite number within the interval $(\lambda\mathbb{E}[X], \bar{\mu}_2)$. In addition, we define $\tilde{\mu}_2 := \sup \mathcal{K}$ and $\hat{\mu}_2 := \inf \mathcal{K}$. We show that $\mathcal{K}$ is nonempty by comparing the order between $h^c(p^*)$ and $h^r(t^*)$. Notice that $\mu_2 = (\mu_1 - \lambda)(1 - \mathbb{E}[X]) + \lambda\mathbb{E}[X]$ is the threshold between $p^* < 1$ and $p^* = 1$. We divide our discussion into two cases: Case (a) is for $\mu_2 \ge (\mu_1 - \lambda)(1 - \mathbb{E}[X]) + \lambda\mathbb{E}[X]$, and Case (b) is for $\mu_2 < (\mu_1 - \lambda)(1 - \mathbb{E}[X]) + \lambda\mathbb{E}[X]$.

(a) When $\mu_2 \ge (\mu_1 - \lambda)(1 - \mathbb{E}[X]) + \lambda\mathbb{E}[X]$, we show that if there ever exists a $\mu_2^0 \in [(\mu_1 - \lambda)(1 - \mathbb{E}[X]) + \lambda\mathbb{E}[X], \bar{\mu}_2)$ at which $h^c(p^*) = h^r(t^*)$, i.e., $\mu_2^0 \in \mathcal{K}$. Then we have $h^c(p^*) > h^r(t^*)$ for all $\mu_2 > \mu_2^0$, and $h^c(p^*) < h^r(t^*)$ for all $(\mu_1 - \lambda)(1 - \mathbb{E}[X]) + \lambda\mathbb{E}[X] \le \mu_2 < \mu_2^0$. Otherwise, $h^c(p^*) > h^r(t^*)$ always holds. Next, in Part (a.1) we show that when $1 - t^* + F(t^*) - t^*(1 - f(t^*)) < 0$, $h^c(p^*) > h^r(t^*)$. In Part (a.2) we show that when $1 - t^* + F(t^*) - t^*(1 - f(t^*)) \ge 0$ and $h^c(p^*) \ge h^r(t^*)$, $h^c(p^*) - h^r(t^*)$ keeps being positive as μ_2 increases.

(a.1) When $1 - t^* + F(t^*) - t^*(1 - f(t^*)) < 0$, we show that $h^c(p^*) > h^r(t^*)$. This proof uses several results that are shown later. By the proof of Proposition 5 and Proposition 6, we know 1) $h^r(t)$ is a strictly convex function of t and $\bar{t}$ is its $\arg\min$; 2) $1 - t^* + F(t^*) - t^*(1 - f(t^*)) < 0 \Rightarrow (1 - f(t^*))\mu_1 - \mu_2 > 0 \Rightarrow \bar{t} < t^* < p^*$, where $t^* < p^*$ is the result proved in the subsection 2b.1. These two results imply that t^* and p^* both lie in the interval on which $h^r(t)$ is increasing. Thus we have $h^r(p^*) > h^r(t^*)$. This result together with Lemma B.4 below implies that $h^c(p^*) \ge h^r(p^*) > h^r(t^*)$.

Lemma B.4 *$h^c(p) \geq h^r(t)$ for all $p = t$.*

The proof of Lemma (B.4) is provided in Section B.6.1.

(a.2) When $1 - t^* + F(t^*) - t^*(1 - f(t^*)) \geq 0$, we show that if at some μ_2, we have $h^c(p^*) \geq h^r(t^*)$, then as μ_2 increases, $h^c(p^*) - h^r(t^*)$ keeps being positive. First, we calculate

$$\begin{aligned} h^c(p^*) - h^r(t^*) &= \frac{p^*}{\mu_1 - \lambda p^*} + \frac{1 - p^* + p^*\mathbb{E}[X]}{\mu_2 - \lambda(1 - p^* + p^*\mathbb{E}[X])} - \left(\frac{t^*}{\mu_1 - \lambda t^*} + \frac{1 - t^* + F(t^*)}{\mu_2 - \lambda(1 - t^* + F(t^*))} \right) \\ &= \frac{1}{\mu_2 - \lambda(1 - p^* + p^*\mathbb{E}[X])} - \frac{1 + F(t^*) - t^* f(t^*)}{\mu_2 - \lambda(1 - t^* + F(t^*))} \\ &= \frac{\mu_2 - \lambda(1 - t^* + F(t^*)) - (1 + F(t^*) - t^* f(t^*))(\mu_2 - \lambda(1 - p^* + p^*\mathbb{E}[X]))}{(\mu_2 - \lambda(1 - p^* + p^*\mathbb{E}[X]))(\mu_2 - \lambda(1 - t^* + F(t^*)))}. \end{aligned} \tag{B.9}$$

Under stability conditions, the denominator of (B.9) is non-negative, so we focus on the sign of the numerator. Let

$$\Delta W(\mu_2) = \mu_2 - \lambda(1 - t^* + F(t^*)) - (1 + F(t^*) - t^* f(t^*))(\mu_2 - \lambda(1 - p^* + p^*\mathbb{E}[X]))$$

denotes the numerator, then

$$\frac{\partial \Delta W(\mu_2)}{\partial \mu_2} = 1 - (1 + F(t^*) - t^* f(t^*)) \left(1 + \lambda(1 - \mathbb{E}[X]) \frac{\partial p^*}{\partial \mu_2} \right) \tag{B.10}$$

$$+ \lambda(1 - f(t^*)) \frac{\partial t^*}{\partial \mu_2} + t^* f'(t^*)(\mu_2 - \lambda(1 - p^* + p^*\mathbb{E}[X])) \frac{\partial t^*}{\partial \mu_2}. \tag{B.11}$$

We show that if $h^c(p^*) \geq h^r(t^*)$, then $\partial \Delta W(\mu_2)/\partial \mu_2 > 0$, i.e., $\Delta W(\mu_2)$ increases in μ_2. To this end, we show (B.10)+(B.11)> 0.

Because $\partial p^*/\partial \mu_2 = -1/(2\lambda(1 - \mathbb{E}[X]))$, and $1 + F(t^*) - t^* f(t^*) < 1$ for all $t^* > 0$, then we have

$$\text{(B.10)} = 1 - \frac{1}{2}(1 + F(t^*) - t^* f(t^*)) > \frac{1}{2}.$$

We show (B.11)$\geq -1/2$. From Equation (B.8), we know $\partial t^*/\partial \mu_2 = -1/(2\lambda(1 - f(t^*)) + (\mu_1 - \lambda t^*) f'(t^*))$. In addition, by Equation (B.3), $\mu_2 - \lambda(1 - p^* + p^*\mathbb{E}[X]) = (\mu_1 - \lambda p^*)(1 - \mathbb{E}[X])$, so that (B.11) is equivalent to

$$-\frac{\lambda(1 - f(t^*)) + t^* f'(t^*)(\mu_1 - \lambda p^*)(1 - \mathbb{E}[X])}{2\lambda(1 - f(t^*)) + (\mu_1 - \lambda t^*) f'(t^*)}. \tag{B.12}$$

Lemma B.5 below shows that, when $h^c(p^*) \geq h^r(t^*)$, term (B.12) $\geq -1/2$.

Lemma B.5 *When $1 - t^* + F(t^*) - t^*(1 - f(t^*)) \geq 0$ and $h^c(p^*) \geq h^r(t^*)$, term* (B.12) $\geq -\frac{1}{2}$.

The proof of Lemma (B.5) is provided in Section B.6.1.

Next, we show that when $\mu_2 = (\mu_1 - \lambda)(1 - \mathbb{E}[X]) + \lambda\mathbb{E}[X]$, it is possible that $h^c(p^*) < h^r(t^*)$. If so, there exists a unique $\mu_2^0 \in [(\mu_1 - \lambda)(1 - \mathbb{E}[X]) + \lambda\mathbb{E}[X], \bar{\mu}_2)$ such that $\mu_2^0 \in \mathcal{K}$, i.e., $h^c(p^*) = h^r(t^*)$. Otherwise, $h^c(p^*) > h^r(t^*)$ always holds.

Assume that at $\mu_2 = (\mu_1 - \lambda)(1 - \mathbb{E}[X]) + \lambda\mathbb{E}[X]$, $h^c(p^*) < h^r(t^*)$, recall in Case 2a.2, we prove that when $\mu_2 = \bar{\mu}_2$, $h^c(p^*) > h^r(t^*)$, thus there must be at least one μ_2^0 under which $h^c(p^*) = h^r(t^*)$, i.e., $\mu_2^0 \in \mathcal{K}$. If such μ_2^0 exists, we next prove that it is unique, by proving that $h^c(p^*) > h^r(t^*)$ for all $\mu_2 > \mu_2^0$. Recall that at μ_2^0, $h^c(p^*) = h^r(t^*)$. As μ_2 increases, if $1 - t^* + F(t^*) - t^*(1 - f(t^*)) < 0$, then as by (a.1) $h^c(p^*) > h^r(t^*)$; if

$1-t^*+F(t^*)-t^*(1-f(t^*))\geq 0$, then by (a.2) the value of $h^c(p^*)-h^r(t^*)$ is positive. Thus $h^c(p^*)>h^r(t^*)$ always holds for all $\mu_2>\mu_2^0$. Assume now that at $\mu_2=(\mu_1-\lambda)(1-\mathbb{E}[X])+\lambda\mathbb{E}[X]$, $h^c(p^*)>h^r(t^*)$, then following the same proof, $h^c(p^*)>h^r(t^*)$ holds for all $(\mu_1-\lambda)(1-\mathbb{E}[X])+\lambda\mathbb{E}[X]\leq\mu_2<\bar{\mu}_2$.

Notice that if μ_2^0 exists, because of its uniqueness, we have $\mu_2^0=\tilde{\mu}_2$. If such μ_2^0 does not exist, we show in Case (b) that when $\mu_2<(\mu_1-\lambda)(1-\mathbb{E}[X])+\lambda\mathbb{E}[X]$, $\tilde{\mu}_2$ exists.

(b) When $\mu_2<(\mu_1-\lambda)(1-\mathbb{E}[X])+\lambda\mathbb{E}[X]$, $p^*=1>t^*$. To simplify the notations, we let $\beta=\mu_2/\lambda$ and $\phi=\mu_1/\lambda$, the average waiting time function is reduced to

$$h^c(p^*)=\frac{1}{\mu_1-\lambda}+\frac{\mathbb{E}[X]}{\mu_2-\lambda\mathbb{E}[X]}=\lambda\left(\frac{1}{\phi-1}+\frac{\mathbb{E}[X]}{\beta-\mathbb{E}[X]}\right), \tag{B.13}$$

and

$$h^r(t^*)=\frac{t^*}{\mu_1-\lambda t^*}+\frac{1-t^*+F(t^*)}{\mu_2-\lambda(1-t^*+F(t^*))}=\lambda\left(\frac{t^*}{\phi-t^*}+\frac{1-t^*+F(t^*)}{\beta-(1-t^*+F(t^*))}\right). \tag{B.14}$$

Consider the difference $\Delta V(\mu_2)=h^c(p^*)-h^r(t^*)$. We prove that when $t^*=0$, $\Delta V(\mu_2)>0$, and when $t^*\to 1$, $\Delta V(\mu_2)<0$. Therefore, there exists at least one $t^*\in(0,1)$ at which $\Delta V(\mu_2)=0$.

By Equation (B.7), we have $\beta=1-t^*+F(t^*)+(\phi-t^*)(1-f(t^*))$. When $t^*=0$,

$$\Delta V(\mu_2)=\lambda\left(\frac{1}{\phi-1}+\frac{\mathbb{E}[X]}{\beta-\mathbb{E}[X]}-\left(\frac{t^*}{\phi-t^*}+\frac{1-t^*+F(t^*)}{\beta-(1-t^*+F(t^*))}\right)\right)=\lambda\left(\frac{1}{\phi-1}+\frac{\mathbb{E}[X]}{\beta-\mathbb{E}[X]}-\frac{1}{\phi}\right)>0.$$

When $t^*\to 1$,

$$\begin{aligned}\Delta V(\mu_2)&=\lambda\left(\frac{1}{\phi-1}+\frac{\mathbb{E}[X]}{\beta-\mathbb{E}[X]}-\left(\frac{t^*}{\phi-t^*}+\frac{1-t^*+F(t^*)}{\beta-(1-t^*+F(t^*))}\right)\right)\\&=\lambda\left(\frac{\mathbb{E}[X]}{1-t^*+F(t^*)+(\phi-t^*)(1-f(t^*))-\mathbb{E}[X]}-\frac{1-t^*+F(t^*)}{(\phi-t^*)(1-f(t^*))}\right)\\&=-\frac{\lambda\mathbb{E}[X]}{2f'(1)(\phi-1)^2}<0,\end{aligned}$$

where the last equality is by Lemma B.6 below. The strictly less than 0 is because, by the definition of f, we must have $f'(1)>0$ and finite.

Lemma B.6 *For fixed $\phi>1$,*

$$\frac{\mathbb{E}[X]}{1-t+F(t)+(\phi-t)(1-f(t))-\mathbb{E}[X]}-\frac{1-t+F(t)}{(\phi-t)(1-f(t))}\to-\frac{\mathbb{E}[X]}{2f'(1)(\phi-1)^2}\ \text{as } t\to 1.$$

The proof of Lemma (B.6) is provided in Section B.6.1.

We then combine this result with the result we obtain in Part (a) to conclude that there exists $\tilde{\mu}_2$ such that $h^c(p^*)\geq h^r(t^*)$ when $\mu_2\geq\tilde{\mu}_2$, and when $\lambda\mathbb{E}[X]<\mu_2<\tilde{\mu}_2$, there exists $\hat{\mu}_2$ such that $h^c(p^*)<h^r(t^*)$ when $\lambda\mathbb{E}[X]<\mu_2<\hat{\mu}_2$.

If at $\mu_2=(\mu_1-\lambda)(1-\mathbb{E}[X])+\lambda\mathbb{E}[X]$, we have $h^c(p^*)\leq h^r(t^*)$, then by the conclusion in Part (a), there exists a unique $\mu_2\geq(\mu_1-\lambda)(1-\mathbb{E}[X])+\lambda\mathbb{E}[X]$ that belongs to the set $\mathcal{K}$ and this μ_2 is $\tilde{\mu}_2$, which can be obtained by solving Equation (B.9) $=0$. In addition, because at $\mu_2\to\lambda\mathbb{E}[X]$, we have $h^c(p^*)<h^r(t^*)$, it is not guaranteed that $\mu_2<(\mu_1-\lambda)(1-\mathbb{E}[X])+\lambda\mathbb{E}[X]$ that belongs to the set $\mathcal{K}$ exists. If there is no such μ_2, we have $\mathcal{K}$ being singleton set only containing $\tilde{\mu}_2$. In this case, we have $h^c(p^*)\geq h^r(t^*)$ when $\mu_2\geq\tilde{\mu}_2$, and vice versa. Otherwise, there might be a set of such $\mu_2\in\mathcal{K}$. Because $\hat{\mu}_2=\inf\mathcal{K}$, we must have $\hat{\mu}_2<\tilde{\mu}_2$, and

$h^c(p^*) \geq h^r(t^*)$ when $\mu_2 \geq \tilde{\mu}_2$, and $h^c(p^*) < h^r(t^*)$ when $\mu_2 < \hat{\mu}_2$. Notice that $\hat{\mu}_2$ must be strictly greater than $\lambda\mathbb{E}[X]$ because $h^r(t^*) - h^c(p^*)$ is strictly greater than 0 when $\mu_2 \to \lambda\mathbb{E}[X]$, both p^* and t^* are continuous in μ_2, and both $h^c(\cdot)$ and $h^r(\cdot)$ are continuous functions.

If at $\mu_2 = (\mu_1 - \lambda)(1 - \mathbb{E}[X]) + \lambda\mathbb{E}[X]$, we still have $h^c(p^*) > h^r(t^*)$, then by the conclusion in Part (a), there exists no $\mu_2 \geq (\mu_1 - \lambda)(1 - \mathbb{E}[X]) + \lambda\mathbb{E}[X]$ that belongs to the set $\mathcal{K}$. However, because at $\mu_2 \to \lambda\mathbb{E}[X]$, we have $h^c(p^*) < h^r(t^*)$, there is a change of sign when μ_2 belongs to the interval $(\lambda\mathbb{E}[X], (\mu_1 - \lambda)(1 - \mathbb{E}[X]) + \lambda\mathbb{E}[X])$. So there must be at least one $\mu_2 \in (\lambda\mathbb{E}[X], (\mu_1 - \lambda)(1 - \mathbb{E}[X]) + \lambda\mathbb{E}[X])$ that belongs to the set $\mathcal{K}$. The exact set $\mathcal{K}$ can be obtained by solving (B.13) = (B.14), and $\tilde{\mu}_2$ and $\hat{\mu}_2$ can be obtained afterwards ($\tilde{\mu}_2$ and $\hat{\mu}_2$ can be equal). Then we have $h^c(p^*) \geq h^r(t^*)$ when $\mu_2 \geq \tilde{\mu}_2$, and $h^c(p^*) < h^r(t^*)$ when $\mu_2 < \hat{\mu}_2$.

We then prove that by fixing f and λ, the change of $\tilde{\mu}_2$ is continuous with respect to the change of μ_1. By the discussion above, $\tilde{\mu}_2$ is obtained either by solving Equation (B.9) = 0 if it has a solution or by solving (B.13) = (B.14) if Equation (B.9) = 0 has no solution. Within each Equation, $\tilde{\mu}_2$ is continuous with respect to μ_1. These two equations coincide when $p^* = 1$, and p^* is continuous with respect to μ_1 as well. Q.E.D.

B.6.1. Auxiliary Results Used in the Proofs of Theorem 1 and Proposition 4

Proof of Lemma B.2: At $t^* = f^{-1}(\mathbb{E}[X])$, we have t^* solves

$$(\mu_1 - \lambda t^*)(1 - \mathbb{E}[X]) = \mu_2 - \lambda(1 - t^* + F(t^*)). \tag{B.15}$$

p^* solves

$$(\mu_1 - \lambda p^*)(1 - \mathbb{E}[X]) = \mu_2 - \lambda(1 - p^* + p^*\mathbb{E}[X]). \tag{B.16}$$

The proof is by contradiction. First, if $p^* = t^*$, then the LHS of (B.15) and (B.16) are the same. But the RHS of (B.15) is greater than the RHS of (B.16) since $p^*\mathbb{E}[X] > F(p^*)$. Contradiction. Second, if $p^* < t^*$, Compared with both sides of Equation (B.15), the LHS of (B.16) increases and the RHS of (B.16) decreases. It is impossible that Equality (B.16) holds. Contradiction. Q.E.D.

Proof of Lemma B.3: When $t^* > f^{-1}(\mathbb{E}[X])$, we have t^* solves

$$(\mu_1 - \lambda t^*)(1 - f(t^*)) = \mu_2 - \lambda(1 - t^* + F(t^*)). \tag{B.17}$$

p^* ($p^* \neq 1$) solves

$$(\mu_1 - \lambda p^*)(1 - \mathbb{E}[X]) = \mu_2 - \lambda(1 - p^* + p^*\mathbb{E}[X]). \tag{B.18}$$

The proof is by contradiction. Suppose $p^* \leq t^*$, then the LHS of (B.17) is strictly less than the LHS of (B.18). But the RHS of (B.17) is strictly greater than the RHS of (B.18) because $\mu_2 - \lambda(1 - p^* + p^*\mathbb{E}[X]) < \mu_2 - \lambda(1 - p^* + F(p^*)) < \mu_2 - \lambda(1 - t^* + F(t^*))$. Thus t^* and p^* could not be the solutions of equation (B.17) and equation (B.18) at the same time. Contradiction. Q.E.D.

Proof of Lemma B.4: We prove this result by showing that $h^c(p) - h^r(t) \geq 0$, $\forall p = t$.

$$h^c(p) - h^r(t) = \frac{1}{\mu_2 - \lambda(1 - p + p\mathbb{E}[X])} - \frac{1 + F(t) - tf(t)}{\mu_2 - \lambda(1 - t + F(t))} \geq 0.$$

where the inequality is because $1 + F(t) - tf(t) \leq 1$ and $1 - p + p\mathbb{E}[X] \geq 1 - t + F(t)$ $\forall p = t$, because $p\mathbb{E}[X] \geq F(t)$. Q.E.D.

Proof of Lemma B.5: When $h^c(p^*) \geq h^r(t^*)$, we have

$$\frac{1}{\mu_2 - \lambda(1-p^*+p^*\mathbb{E}[X])} \geq \frac{1+F(t^*)-t^*f(t^*)}{\mu_2-\lambda(1-t^*+F(t^*))},$$

which is equivalent to

$$\begin{aligned}\mu_2 - \lambda(1-p^*+p^*\mathbb{E}[X]) &\leq \frac{\mu_2-\lambda(1-t^*+F(t^*))}{1+F(t^*)-t^*f(t^*)}\\ \Rightarrow (\mu_1-\lambda p^*)(1-\mathbb{E}[X]) &\leq \frac{(\mu_1-\lambda t^*)(1-f(t^*))}{1+F(t^*)-t^*f(t^*)} \text{ by the definitions of } p^* \text{ and } t^*\\ &\leq \frac{(\mu_1-\lambda t^*)(1-f(t^*))}{2t^*(1-f(t^*))} \Leftrightarrow (\mu_1-\lambda p^*)(1-\mathbb{E}[X]) \leq \frac{(\mu_1-\lambda t^*)(1-f(t^*))}{2t^*(1-f(t^*))}.\end{aligned}$$

The third inequality is because $1+F(t^*)-t^*f(t^*) = 1-t^*+F(t^*)+t^*(1-f(t^*)) \geq 2t^*(1-f(t^*))$ by assumption. Rearranging the last inequality, we get $t^*(\mu_1-\lambda p^*)(1-\mathbb{E}[X]) \leq (\mu_1-\lambda t^*)/2 \Leftrightarrow t^*f'(t^*)(\mu_1-\lambda p^*)(1-\mathbb{E}[X]) \leq (\mu_1-\lambda t^*)f'(t^*)/2 \Rightarrow$ (B.12) $\geq -1/2$. Q.E.D.

Proof of Lemma B.6: Let $\omega(t) = 1-t+F(t)$ and $\sigma(t) = (\phi-t)(1-f(t))$. Then we have $\omega'(t) = -(1-f(t))$, $\omega''(t) = f'(t)$, $\sigma'(t) = -(1-f(t))-(\phi-t)f'(t)$, and $\sigma''(t) = 2f'(t)-(\phi-t)f''(t)$. As $t \to 1$, we have $\omega(1) = \mathbb{E}[X]$, $\omega'(1) = 0$, $\omega''(1) = f'(1)$, $\sigma(1) = 0$, $\sigma'(1) = -(\phi-1)f'(1)$, and $\sigma''(1) = 2f'(1)-(\phi-1)f''(1)$. Thus we get

$$\frac{\mathbb{E}[X]}{1-t+F(t)+(\phi-t)(1-f(t))-\mathbb{E}[X]} - \frac{1-t+F(t)}{(\phi-t)(1-f(t))} = -\frac{\omega(\omega+\sigma-\mathbb{E}[X])-\sigma\mathbb{E}[X]}{\sigma(\omega+\sigma-\mathbb{E}[X])}. \tag{B.19}$$

Apply L'Hospital's rule twice on (B.19) and plug in those quantities, we get

$$-\frac{\omega''(\omega+\sigma-\mathbb{E}[X])+2\omega'(\omega'+\sigma')+\omega(\omega''+\sigma'')-\sigma''\mathbb{E}[X]}{\sigma''(\omega+\sigma-\mathbb{E}[X])+2\sigma'(\sigma'+\omega')+\sigma(\omega''+\sigma'')} = -\frac{\omega\omega''}{2\sigma'^2} = -\frac{\mathbb{E}[X]}{2f'(1)(\phi-1)^2},$$

at $t = 1$. Q.E.D.

B.7. Proof of Proposition 5

Proof: To show that the first best is of threshold type, we first express the explicit formula for the average steady-state waiting time $\mathbb{E}[W(g,t)]$ under routing strategy g below:

$$\mathbb{E}[W(g,t)] = \frac{p}{\mu_1-\lambda p} + \frac{1-p+D(p)}{\mu_2-\lambda(1-p+D(p))},$$

where $p = \int_0^1 g(x)dx$ and $D(p) = \int_0^1 g(x)f(x)dx$. We show that for any g that leads to the same realization of p, the one-threshold type of strategy always outperforms or achieves comparable results. Recall the average waiting time under threshold strategy t has been defined as

$$h^{fb}(t) = \frac{t}{\mu_1-\lambda t} + \frac{1-t+F(t)}{\mu_2-\lambda(1-t+F(t))}.$$

Fixing g, we show $h^{fb}(p) \leq \mathbb{E}[W(g,t)]$. It is enough to show

$$\frac{1-p+F(p)}{\mu_2-\lambda(1-p+F(p))} \leq \frac{1-p+D(p)}{\mu_2-\lambda(1-p+D(p))},$$

which is equivalent to show $F(p) \leq D(p)$. Because f is strictly increasing, we have $D(p) = \int_0^1 g(x)f(x)dx \geq \int_0^1 \mathbb{1}_{x\leq p} f(x)dx = \int_0^p f(x)dx = F(p)$. Thus the result holds.

To find the first solution, we need to find the strategy $t \in [0,1]$ that minimizes $h^{fb}(t)$. To do so, we first prove that $h^{fb}(t)$ is strictly convex in t. Under the refined stability conditions, the derivative of the first term of $h^{fb}(t)$ is $\mu_1/(\mu_1-\lambda t)^2>0$, and the second derivative of it is $2\lambda\mu_1(\mu_1-\lambda t)/(\mu_1-\lambda t)^4>0$. So the first term is strictly convex. The derivative of the second term of $h^{fb}(t)$ is $(f(t)-1)\mu_2/(\mu_2-\lambda(1-t+F(t)))^2\le 0$, and the second derivative of it is

$$\frac{\mu_2(\mu_2-\lambda(1-t+F(t)))(f'(t)(\mu_2-\lambda(1-t+F(t)))+2\lambda(f(t)-1)^2)}{(\mu_2-\lambda(1-t+F(t)))^4}>0$$

since $f'(t)>0$, thus the second term is also strictly convex. By the property of the strictly convex functions, the function $h^{fb}(t)$ is also strictly convex. Thus a minimizer $\bar{t}$ that minimizes $h^{fb}(t)$ always exists and is unique.

The derivative of $h^{fb}(t)$ is

$$h^{fb'}(t)=\frac{\mu_1}{(\mu_1-\lambda t)^2}+\frac{(f(t)-1)\mu_2}{(\mu_2-\lambda(1-t+F(t)))^2}. \tag{B.20}$$

Notice that the minimizer $\bar{t}$ should be strictly less than the upper bound of t that satisfies the refined stability condition, i.e., $\bar{t}<\min(\mu_1/\lambda,1)$, because $h^{fb'}(\mu_1/\lambda)=\infty$ and $h^{fb'}(1)=\frac{\mu_1}{(\mu_1-\lambda)^2}>0$.

Case 1. When $\lambda\le\mu_2-\sqrt{\mu_1\mu_2}$, $\bar{t}=0$ is the first best solution. This is because by the strict convexity property, $t=0$ is the minimizer if and only if $h^{fb'}(0)\ge 0\Leftrightarrow 1/\mu_1-\mu_2/(\mu_2-\lambda)^2\ge 0\Leftrightarrow\lambda\le\mu_2-\sqrt{\mu_1\mu_2}$.

Case 2. When $\lambda>\mu_2-\sqrt{\mu_1\mu_2}$, we have $h^{fb'}(0)<0$. In addition, because $f(t)-1<0$ at $t=t_s$, where t_s is the solution of $\mu_2-\lambda(1-t+F(t))=0$, we have $h^{fb'}(t_s)=-\infty$. Thus at the lower bound of t that satisfies the refined stability condition, we always have $h^{fb'}<0$, and at the upper bound of t, we always have $h^{fb'}>0$. So there must be a $\bar{t}\in(0,1)$ that uniquely solves Equation (B.20) $=0$ and $\bar{t}$ is the first best solution. Q.E.D.

B.8. Proof of Proposition 6

Proof: **Case 1**. When $\mu_1\le\mu_2$,

1.1 By Proposition 2 and Proposition 5, when $0<\lambda\le\mu_2-\sqrt{\mu_1\mu_2}$, we always have $t^*=\bar{t}=0$. When $\mu_2-\sqrt{\mu_1\mu_2}<\lambda\le\mu_2-\mu_1$, we always have $\bar{t}>t^*=0$.

1.2 Otherwise, when $\lambda>\mu_2-\sqrt{\mu_1\mu_2}$, then $t^*>0$ and $\bar{t}>0$, plugging t^* from Equation (B.7) into (B.20), we get

$$\begin{aligned}\frac{\mu_1}{(\mu_1-\lambda t^*)^2}+\frac{(f(t^*)-1)\mu_2}{(\mu_2-(1-t^*+F(t^*))\lambda)^2}&=\frac{\mu_1}{(\mu_1-\lambda t^*)^2}+\frac{(f(t^*)-1)\mu_2}{(1-f(t^*))^2(\mu_1-\lambda t^*)^2}\\&=\frac{(1-f(t^*))\mu_1-\mu_2}{(1-f(t^*))(\mu_1-\lambda t^*)^2}.\end{aligned} \tag{B.21}$$

Because h^{fb} is strictly, and $\bar{t}>0$ is the minimizer of h^{fb}, if at $t=t^*$ its derivative (B.21) ≤ 0, we must have $t^*\le\bar{t}$; otherwise, we must have $t^*>\bar{t}$. However, when $\mu_1\le\mu_2$, we always have $(1-f(t^*))\mu_1-\mu_2\le 0$, which implies (B.21) ≤ 0, so we conclude that $t^*\le\bar{t}$.

Case 2. When $\mu_1>\mu_2$, Lemma B.7 gives the comparison between t^* and $\bar{t}$ based on the sign of (B.21).

Lemma B.7 *If $(1-f(t^*))\mu_1<\mu_2$, $t^*<\bar{t}$; if $(1-f(t^*))\mu_1>\mu_2$, $t^*>\bar{t}$; if $(1-f(t^*))\mu_1=\mu_2$, $t^*=\bar{t}$.*

The proof of Lemma B.7 is simple and we skip it.

We then refine the criterion provided in Lemma B.7 to get rid of t^*. To do that, we first analyze the sensitivity of t^* with respect to λ. We rewrite Equation (B.7) as

$$\mu_2-(1-f(t^*))\mu_1=(1-t^*+F(t^*)-(1-f(t^*)t^*)\,\lambda. \tag{B.22}$$

Take the derivative of (B.22) with respect to λ on both sides, we get

$$((\mu_1-\lambda t^*)f'(t^*)+2\lambda(1-f(t^*)))\,\frac{\partial t^*}{\partial\lambda}=1-t^*+F(t^*)-(1-f(t^*))t^*.$$

Note $(\mu_1-\lambda t^*)f'(t^*)+2\lambda(1-f(t^*))>0$. If $1-t^*+F(t^*)>(1-f(t^*))t^*$, we have $\partial t^*/\partial\lambda>0$. If $1-t^*+F(t^*)<(1-f(t^*))t^*$, we have $\partial t^*/\partial\lambda<0$. If $1-t^*+F(t^*)=(1-f(t^*))t^*$, we have $\partial t^*/\partial\lambda=0$.

When $\lambda\to 0$, according to Equation (B.22), $\hat{t}$ that solves $\mu_2-(1-f(\hat{t}))\mu_1=0$ is the refined patient equilibrium. If at $t=\hat{t}$, $1-\hat{t}+F(\hat{t})-(1-f(\hat{t}))\hat{t}<0$, then by the sensitivity analysis above, $\partial\hat{t}/\partial\lambda<0$, $\hat{t}$ decreases with respect to λ. As λ increases, $\hat{t}$ decreases, and thus the LHS of Equation (B.22) decreases and its value becomes strictly less than 0. So the new equilibrium that solves (B.22) still satisfies $1-t^*+F(t^*)-(1-f(t^*)t^*<0$. This implies that as λ increases, we always have t^* decreases, and thus $\mu_2-(1-f(t^*))\mu_1<\mu_2-(1-f(\hat{t}))\mu_1=0$ always holds. By Lemma B.7, we have $t^*>\bar{t}$.

The same type of argument can be applied to the case when $1-\hat{t}+F(\hat{t})-(1-f(\hat{t}))\hat{t}>0$.

If at $t=\hat{t}$, $1-\hat{t}+F(\hat{t})-(1-f(\hat{t}))\hat{t}=0$, then we have $\partial\hat{t}/\partial\lambda=0$. This implies the change of λ does not change the equilibrium. This trend continues as λ increases. So by Lemma B.7, we always have $t^*=\bar{t}$ regardless of λ. Q.E.D.

B.9. Proof of Theorem 2

Proof: When $(\lambda,\mu_1,\mu_2,f)\in\mathcal{S}^{refined}$, we first show that patients' equilibrium strategy under priority rule $q\in(0,1]$ is of threshold type. To show that, suppose all patients apply a mixed strategy in choosing between the two services. Let $g:[0,1]\to[0,1]$ denote patients' strategy and let $p=\int_0^1 g(x)dx$. In addition, suppose there are two classes, and the expected waiting time of class 1 is denoted by T_1 and the expected waiting time of class 2 is denoted by T_2. By definition, because patients are indifferent between the queues, the expected waiting times at both queues should be the same for all types of patients. We discuss two cases. First, under a policy that prioritizes the follow-up patients with probability q over the other patients at the in-person service, the mixed strategy equilibrium for all types should satisfy

$$\frac{1}{\mu_1-\lambda p}+f(t)(qT_1(p)+(1-q)T_2(p))=T_2(p),\forall t\in[0,1]. \tag{B.23}$$

The LHS of Equation (B.23) is the expected waiting time at the telemedicine queue and the RHS of Equation (B.23) is the expected waiting time at the in-person queue.

For fixed strategy g, the LHS is strictly increasing while the RHS is a constant, thus if the two lines intersect, there is at most one intersection, which we denote as $t(q)$.

Case 1. If $t(q)\in[0,1]$ exists under strategy g, then because f is strictly increasing, we have LHS<RHS for all $t<t(q)$, and LHS>RHS for all $t>t(q)$. That is, all types $t<t(q)$ deviate to the telemedicine service and all types $t>t(q)$ deviate to the in-person service. Thus $g(t)=1$ for $t\le t(q)$, $g(t)=0$ for $t>t(q)$, and $p=t(q)$ is the equilibrium.

Case 2. If LHS<RHS for all t under strategy g, then we define $t(q)=1$, and $g(t)=1$ for $t\le t(q)$ is the equilibrium.

Case 3. If LHS>RHS for all t under strategy g, then we define $t(q)=0$, and $g(t)=0$ for $t\ge t(q)$ is the equilibrium.

The same type of proof can be applied to the case of prioritizing with probability q the in-person patients in the in-person queue over the follow-up patients. Then, Equation (B.23) is replaced by

$$\frac{1}{\mu_1 - \lambda p} + f(t)T_2(p) = qT_1(p) + (1-q)T_2(p),$$

where the LHS is the expected waiting time at the telemedicine queue and the RHS is the expected waiting time at the in-person queue.

We then prove that the one-threshold type of equilibrium always exists under priority rule $q \in (0,1]$, and we derive the optimal q^* that can induce the first best solution.

Suppose the service intensity for class 1 patient is ρ_1 and the service intensity for class 2 patient is ρ_2, under the assumption of $M/M/1$ queue, the expected waiting time of each class is:

$$T_1 = \frac{1+\rho_2}{\mu_2(1-\rho_1)},$$

$$T_2 = \frac{1-\rho_1(1-\rho_1-\rho_2)}{\mu_2(1-\rho_1)(1-\rho_1-\rho_2)}.$$

The analysis can be divided into two cases.

Case 1. When $t^* < \bar{t}$, we prioritize the follow-up patients with probability $q \in (0,1]$ over the other patients in the in-person service. Then the service intensity for the class 1 patients equals $\rho_1 = \lambda q F(t)/\mu_2$, and the service intensity for the class 2 patients equals $\rho_2 = \lambda(1-t+(1-q)F(t))/\mu_2$, where t is the threshold. For a given q, the Nash equilibrium threshold $t(q) \in (0,1)$ should satisfy:

$$\frac{1}{\mu_1 - \lambda t} + f(t)(qT_1 + (1-q)T_2) = T_2. \tag{B.24}$$

Rearranging Equation (B.24), we get

$$q = \frac{\mu_2^2(\mu_2 - \lambda(1-t+F(t) - (1-f(t))(\mu_1 - \lambda t)))}{\lambda\left(\mu_2 F(t)(\mu_2 - \lambda(1-t+F(t) - (1-f(t))(\mu_1-\lambda t))) + \lambda(1-t+F(t))((1-t)f(t)+F(t))(\mu_1-\lambda t)\right)}. \tag{B.25}$$

1.1 When $0 = t^* < \bar{t}$, i.e., $\lambda + \sqrt{\mu_1\mu_2} > \mu_2 \geq \lambda + \mu_1$, we prove that the priority rule could not improve the patient's equilibrium. Equivalently, we prove Equation (B.24) has no solution $t(q) \in (0,1)$, i.e., $t(q) = 0$ is the equilibrium.

As before, we denote $\beta = \mu_2/\lambda$ and $\phi = \mu_1/\lambda$. Then, Equation (B.25) reduces to

$$q = \frac{\beta^2(\beta - (1-t+F(t)) - (1-f(t))(\phi - t))}{\beta F(t)(\beta - (1-t+F(t)) - (1-f(t))(\phi-t)) + (1-t+F(t))((1-t)f(t)+F(t))(\phi-t)}. \tag{B.26}$$

Let

$$\alpha = \frac{1}{q} = \frac{F(t)}{\beta} + \frac{(1-t+F(t))((1-t)f(t)+F(t))(\phi-t)}{\beta^2(\beta-(1-t+F(t))-(1-f(t))(\phi-t))}, \quad \alpha \geq 1. \tag{B.27}$$

We show that when $\beta \geq \phi + 1$, i.e., $\mu_2 \geq \lambda + \mu_1$, Equation (B.27) has no solution $t \in [0,1]$ for all $\alpha \geq 1$. We prove it by showing that (B.27)< 1, which is equivalent to

$$\frac{(1-t+F(t)))(\phi-t)}{\beta-(1-t+F(t))-(1-f(t))(\phi-t)} < \frac{\beta(\beta-F(t))}{(1-t)f(t)+F(t)}, \tag{B.28}$$

for all $t \in [0,1]$. Since $\beta \geq \phi+1$, we have the denominators $\beta - (1-t+F(t)) - (1-f(t))(\phi-t) \geq (1-t)f(t) + F(t)$ because $\beta - (1-t+F(t)) - (1-f(t))(\phi-t) \geq \phi + 1 - (1-t+F(t)) - (1-f(t))(\phi-t) \geq (1-t)f(t) + F(t)$, where the second inequality is because $\phi+1-(1-t+F(t))-(1-f(t))(\phi-t)-(1-t)f(t)-F(t)$ increases in t, and at $t=0$, its value is 0. In addition, we have the numerators $(1-t+F(t)))(\phi-t)<(\phi+1)(\phi+1-F(t))\leq\beta(\beta-F(t))$, where the first inequality is because $1-t+F(t)<\phi+1$ and $\phi-t<\phi+1-F(t)$. Therefore, the inequality holds, and thus, there is no $q\in(0,1]$ that satisfies (B.26) for all $t\in[0,1]$. In this case, the LHS of (B.24)>the RHS of (B.24), meaning the expected waiting time at the telemedicine queue is always greater than the at the in-person queue. Thus, $t(q)=0$ is the equilibrium for all $\in(0,1]$.

1.2 When $0<t^*<\bar{t}$, first, we denote $\tilde{q}$ by plugging $t=\bar{t}$ in Equation (B.25). Then,

$$\tilde{q}=\frac{\mu_2^2(\mu_2-\lambda(1-\bar{t}+F(\bar{t})-(1-f(\bar{t}))(\mu_1-\lambda\bar{t})))}{\lambda\left(\mu_2F(\bar{t})(\mu_2-\lambda(1-\bar{t}+F(\bar{t})-(1-f(\bar{t}))(\mu_1-\lambda\bar{t})))+\lambda(1-\bar{t}+F(\bar{t}))((1-\bar{t})f(\bar{t})+F(\bar{t}))(\mu_1-\lambda\bar{t})\right)}. \tag{B.29}$$

With the definition of $\tilde{q}$, Lemma B.8 below proves that there always exists a threshold type of equilibrium, denoted by $t^*(q)$, which is increasing in q. At $q=0$, we have $t^*(0)=t^*$ (the refined patient equilibrium), and at $q=\tilde{q}$, it can be verified that we have $t^*(\tilde{q})=\bar{t}$ (the system's first best), which implies that $\tilde{q}$ induces the first best solution.

Lemma B.8 *For each $0\leq q<\tilde{q}$, there exists $t^*(q)$ solving Equation* (B.24) *such that $t^*(q)\in[t^*,\bar{t})$, and $t^*(q)$ increases with respect to q.*

The proof of Lemma (B.8) is provided in Section B.9.1.

We then derive the conditions under which a valid $\tilde{q}$ that induces the first best solution exists, i.e., $0<\tilde{q}\leq1$. Recall by (B.7), we have $1/(\mu_1-\lambda t^*)=(1-f(t^*))/(\mu_2-\lambda(1-t^*+F(t^*)))$, where the LHS is increasing in t^* and the RHS is decreasing in t^*. Because $t^*<\bar{t}$, replacing t^* with $\bar{t}$ on both sides, we get LHS>RHS, i.e., $1/(\mu_1-\lambda\bar{t})>(1-f(\bar{t}))/(\mu_2-\lambda(1-\bar{t}+F(\bar{t})))$. Rearranging it, we get $\mu_2-\lambda(1-\bar{t}+F(\bar{t}))-(1-f(\bar{t}))(\mu_1-\lambda\bar{t})>0$. Thus both the denominator and numerator of $\tilde{q}$ are positive. To ensure a valid $\tilde{q}$ exists, the denominator of $\tilde{q}$ needs to be greater than (or equal to) the numerator, as shown below.

$$\begin{aligned}
&\lambda\left(\mu_2F(\bar{t})(\mu_2-\lambda(1-\bar{t}+F(\bar{t})-(1-f(\bar{t}))(\mu_1-\lambda\bar{t})))+\lambda(1-\bar{t}+F(\bar{t}))((1-\bar{t})f(\bar{t})+F(\bar{t}))(\mu_1-\lambda\bar{t})\right)\\
\geq&\mu_2^2(\mu_2-\lambda(1-\bar{t}+F(\bar{t})-(1-f(\bar{t}))(\mu_1-\lambda\bar{t})))\\
\Leftrightarrow&\lambda^2(1-\bar{t}+F(\bar{t}))((1-\bar{t})f(\bar{t})+F(\bar{t}))(\mu_1-\lambda\bar{t})\\
\geq&\mu_2(\mu_2-\lambda F(\bar{t}))(\mu_2-\lambda(1-\bar{t}+F(\bar{t})-(1-f(\bar{t}))(\mu_1-\lambda\bar{t}))).
\end{aligned}$$

The above condition ensures that there exists a valid $0<\tilde{q}\leq1$ at which $\bar{t}$ is a solution to Equation (B.24), i.e., the system can be coordinated by $\tilde{q}$, and $q^*=\tilde{q}$. Otherwise, when $\tilde{q}>1$ is needed to coordinate the system, the monotonicity result in Lemma B.8 implies that $q^*=1$.

Case 2. When $t^*>\bar{t}$, we prioritize the in-person patients with probability $q\in(0,1]$ over the follow-up patients at the in-person service. Then the service intensity for class 1 patients equals $\rho_1=\lambda q(1-t)/\mu_2$, and the service intensity for class 2 patients equals $\rho_2=\lambda((1-q)(1-t)+F(t))/\mu_2$, where t is the threshold. For a given q, the Nash equilibrium threshold strategy $t(q)\in(0,1)$ should satisfy:

$$\frac{1}{\mu_1-\lambda t}+f(t)T_2=qT_1+(1-q)T_2. \tag{B.30}$$

Rearranging Equation (B.30), we get

$$q = \frac{\mu_2^2(\mu_2 - \lambda(1-t+F(t)) - (1-f(t))(\mu_1 - \lambda t))}{\lambda\,(\mu_2(1-t)(\mu_2 - \lambda(1-t+F(t)) - (1-f(t))(\mu_1 - \lambda t)) - \lambda(1-t+F(t))((1-t)f(t)+F(t))(\mu_1 - \lambda t))}. \tag{B.31}$$

Following the same idea, we define

$$\tilde{q} = \frac{\mu_2^2(\mu_2 - \lambda(1-\bar{t}+F(\bar{t})) - (1-f(\bar{t}))(\mu_1 - \lambda \bar{t}))}{\lambda\,(\mu_2(1-\bar{t})(\mu_2 - \lambda(1-\bar{t}+F(\bar{t})) - (1-f(\bar{t}))(\mu_1 - \lambda \bar{t})) - \lambda(1-\bar{t}+F(\bar{t}))((1-\bar{t})f(\bar{t})+F(\bar{t}))(\mu_1 - \lambda \bar{t}))}. \tag{B.32}$$

Again, $\bar{t}$ is the equilibrium solution under priority rule $q = \tilde{q}$. In addition, we prove the following Lemma.

Lemma B.9 *For each* $0 \le q < \tilde{q}$*, there exists* $t^*(q)$ *solving Equation* (B.30) *such that* $t^*(q) \in (\bar{t}, t^*]$*, and* $t^*(q)$ *decreases with respect to* q*.*

The proof of Lemma (B.9) is provided in Section B.9.1.

We then derive the conditions under which a valid $\tilde{q}$ that induces the first best solution exists, i.e., $0 < \tilde{q} \le 1$. Because now $t^* > \bar{t}$, by (B.7), we get $\mu_2 - \lambda(1-\bar{t}+F(\bar{t})) - (1-f(\bar{t}))(\mu_1 - \lambda\bar{t}) < 0$. Thus both the denominator and numerator of $\tilde{q}$ are negative. To ensure a valid $\tilde{q}$ exists, the denominator of $\tilde{q}$ needs to be less than (or equal to) the numerator, as shown below.

$$\begin{aligned}
&\lambda\,(\mu_2(1-\bar{t})(\mu_2 - \lambda(1-\bar{t}+F(\bar{t}) - (1-f(\bar{t}))(\mu_1 - \lambda\bar{t}))) - \lambda(1-\bar{t}+F(\bar{t}))((1-\bar{t})f(\bar{t})+F(\bar{t}))(\mu_1 - \lambda\bar{t}))\\
\le&\mu_2^2(\mu_2 - \lambda(1-\bar{t}+F(\bar{t}) - (1-f(\bar{t}))(\mu_1 - \lambda\bar{t})))\\
\Leftrightarrow&\lambda^2(1-\bar{t}+F(\bar{t}))((1-\bar{t})f(\bar{t})+F(\bar{t}))(\mu_1 - \lambda\bar{t})\\
\ge&\mu_2(\lambda(1-\bar{t}) - \mu_2)(\mu_2 - \lambda(1-\bar{t}+F(\bar{t})) - (1-f(\bar{t}))(\mu_1 - \lambda\bar{t})).
\end{aligned}$$

The above condition ensures that there exists a valid $0 < \tilde{q} \le 1$ at which $\bar{t}$ is a solution to Equation (B.30), i.e., the system can be coordinated by $\tilde{q}$, and $q^* = \tilde{q}$. Otherwise, when $\tilde{q} > 1$ is needed to coordinate the system, the monotonicity result in Lemma B.9 implies that $q^* = 1$. Q.E.D.

B.9.1. Auxiliary Results Used in the Proofs of Theorem 2

PROOF OF LEMMA B.8: By (B.6), t^* has to satisfy

$$\frac{\mu_2 - \lambda(1-t+F(t))}{\mu_1 - \lambda t} = 1 - f(t), \tag{B.33}$$

and by (B.20) $= 0$, $\bar{t}$ has to satisfy

$$\frac{\mu_2 - \lambda(1-t+F(t))}{\mu_1 - \lambda t} = \sqrt{\frac{(1-f(t))\mu_2}{\mu_1}}. \tag{B.34}$$

The LHS of (B.33) and (B.34) are obviously decreasing with respect to t.

With the priority rule, (B.25) can be reduced to

$$\frac{\mu_2 - \lambda(1-t+F(t))}{\mu_1 - \lambda t} = 1 - f(t) + \frac{q\lambda^2(1-t+F(t))((1-t)f(t)+F(t))}{\mu_2(\mu_2 - q\lambda F(t))}. \tag{B.35}$$

Notice that the LHS of Equation (B.33)(B.34)(B.35) are the same, we define it as

$$B(t) := \frac{\mu_2 - \lambda(1-t+F(t))}{\mu_1 - \lambda t}.$$

Take the derivative of $B(t)$ with respect to t, we get

$$\frac{\partial B(t)}{\partial t} = \frac{\lambda(1-f(t))(\mu_1-\lambda t)+\lambda(\mu_2-\lambda(1-t+F(t)))}{(\mu_1-\lambda t)^2} > 0.$$

We then show that if $0 \leq q < \tilde{q}$, we always have a $t^*(q) \in [t^*, \bar{t})$ exists and it increases in q. That is, we show for two arbitrary priority parameters q_1 and q_2 with $0 < q_1 < q_2$, we have $t^*(q_1), t^*(q_2)$ both exist, and $t^*(q_1) < t^*(q_2)$. The proof is by induction.

First, at $q = 0$, obviously, we have $t^*(q) = t^*$. In addition, notice that the RHS of (B.35) subtract the RHS of Equation (B.33) equals $(q\lambda^2(1-t+F(t))((1-t)f(t)+F(t)))/(\mu_2(\mu_2-q\lambda F(t))) > 0$. Thus we have the RHS of (B.35) greater than the RHS of Equation (B.33) for all $t \in (0,1]$ satisfying the refined stability condition. As q increases a little bit from 0, the new solution $t^*(q)$ should exist and be strictly greater than t^*. This is because

$$\frac{q\lambda^2(1-t+F(t))((1-t)f(t)+F(t))}{\mu_2(\mu_2-q\lambda F(t))} = \frac{\lambda^2(1-t+F(t))((1-t)f(t)+F(t))}{\mu_2\left(\frac{\mu_2}{q}-\lambda F(t)\right)},$$

the denominator of the RHS goes to infinity when q is very small. Thus the RHS of (B.35) and the RHS of Equation (B.33) do not differ much when q is close to 0. In addition, since $B(t)$ is increasing in t and the RHS of (B.33) is strictly decreasing in t, Equation (B.35) must have a solution and the solution must be strictly greater than t^*.

Suppose at arbitrary $q_1 \in (0, \tilde{q})$, we have that $t^*(q_1) \in [t^*, \bar{t})$ exists and $t^*(q_1) > t^*$. We show that at $q_2 \in (q_1, \tilde{q})$, there exists $t^*(q_2) \in [t^*, \bar{t})$ such that $t^*(q_2) > t^*(q_1)$.

(a) First, we show that under $q = q_2$, at the point $t = t^*(q_1)$, we have the RHS of Equation (B.35)>LHS of Equation (B.35). At $t = t^*(q_1)$, we have

$$(B(t)-(1-f(t)))\mu_2(\mu_2-q_1\lambda F(t)) > (B(t)-(1-f(t)))\mu_2(\mu_2-q_2\lambda F(t)). \tag{B.36}$$

because $B(t)-(1-f(t)) > 0$ when $t > t^*$ and $q_1 < q_2$. So that at $t = t^*(q_1)$, we have

$$\begin{aligned} q_2\lambda^2(1-t+F(t))((1-t)f(t)+F(t)) &> q_1\lambda^2(1-t+F(t))((1-t)f(t)+F(t)) \text{ because } q_2 > q_1 \\ &= (B(t)-(1-f(t)))\mu_2(\mu_2-q_1\lambda F(t)) \text{ by the definition of } t^*(q_1) \\ &> (B(t)-(1-f(t)))\mu_2(\mu_2-q_2\lambda F(t)) \text{ by inequality (B.36)} \\ \Rightarrow B(t)\mu_2(\mu_2-q_2\lambda F(t)) &< \mu_2(\mu_2-q_2\lambda F(t))(1-f(t)) + q_2\lambda^2(1-t+F(t))((1-t)f(t)+F(t)). \end{aligned}$$

Thus at $t = t^*(q_1)$ and $q = q_2$, we have the RHS of Equation (B.35)>LHS of Equation (B.35).

(b) Next, we show that under $q = q_2$, at the point $t = \bar{t}$, we have the RHS of Equation (B.35)<LHS of Equation (B.35). This is because at $q = \tilde{q}$ and $t = \bar{t}$, by definition of $\tilde{q}$, we have the RHS of Equation (B.35)=LHS of Equation (B.35) because Equation (B.25) and Equation (B.35) are equivalent. Because $q_2 < \tilde{q}$, by rearranging Equation (B.25) at $t = \bar{t}$, we get the RHS of Equation (B.35)<LHS of Equation (B.35) at $q = q_2$.

Combining the results in Part (a) and Part (b), it must be true that there exists $t^*(q_2) \in (t^*(q_1), \bar{t})$ that solves Equation (B.35). Thus the induction result holds. Q.E.D.

Proof of Lemma B.9: Following the same procedure as in the proof of Lemma B.8, with the priority rule, (B.31) can be reduced to

$$\frac{\mu_2-\lambda(1-t+F(t))}{\mu_1-\lambda t}=1-f(t)-\frac{q\lambda^2(1-t+F(t))((1-t)f(t)+F(t))}{\mu_2(\mu_2-q\lambda(1-t))}. \tag{B.37}$$

We then show that if $0\le q<\tilde{q}$, we always have a $t^*(q)\in(\bar{t},t^*]$ exists and it decreases in q. That is, we show for two arbitrary priority parameters q_1 and q_2 with $0<q_1<q_2$, we have $t^*(q_1),t^*(q_2)$ both exist, and $t^*(q_1)>t^*(q_2)$. The proof is by induction.

First, notice that at $q=0$, we have $t^*(q)=t^*$. In addition, notice that the RHS of (B.37) subtract the RHS of Equation (B.33) equals $-(q\lambda^2(1-t+F(t))((1-t)f(t)+F(t)))/(\mu_2(\mu_2-q\lambda(1-t)))<0$. Thus we have the RHS of (B.37) less than the RHS of Equation (B.33) for all $t\in(0,1]$ satisfying the refined stability condition. As q increases a little bit from 0, the new solution $t^*(q)$ should exist and be strictly less than t^*. This is because

$$-\frac{q\lambda^2(1-t+F(t))((1-t)f(t)+F(t))}{\mu_2(\mu_2-q\lambda(1-t))}=-\frac{\lambda^2(1-t+F(t))((1-t)f(t)+F(t))}{\mu_2\left(\frac{\mu_2}{q}-\lambda(1-t)\right)},$$

the denominator of the RHS goes to infinity when q is very small. Thus the RHS of (B.37) and the RHS of Equation (B.33) do not differ much when q is close to 0. In addition, since $B(t)$ is increasing in t and the RHS of (B.33) is strictly decreasing in t, Equation (B.37) must have a solution and the solution must be strictly less than t^*.

Suppose at arbitrary $q_1\in(0,\tilde{q})$, we have that $t^*(q_1)\in(\bar{t},t^*]$ exists and $t^*(q_1)<t^*$. We show that at $q_2\in(q_1,\tilde{q})$, there exists $t^*(q_2)\in(\bar{t},t^*]$ such that $t^*(q_2)<t^*(q_1)$.

(a) First, we show that under $q=q_2$, at the point $t=t^*(q_1)$, we have the RHS of Equation (B.37)<LHS of Equation (B.37). At $t=t^*(q_1)<t^*$, we have

$$(B(t)-(1-f(t)))\mu_2(\mu_2-q_1\lambda(1-t))<(B(t)-(1-f(t)))\mu_2(\mu_2-q_2\lambda(1-t)). \tag{B.38}$$

because $(B(t)-(1-f(t)))<0$ when $t<t^*$. So that at $t=t^*(q_1)$, we have

$$\begin{aligned}-q_2\lambda^2(1-t+F(t))((1-t)f(t)+F(t))&<-q_1\lambda^2(1-t+F(t))((1-t)f(t)+F(t)) \text{ because } q_2>q_1\\ &=(B(t)-(1-f(t)))\mu_2(\mu_2-q_1\lambda(1-t)) \text{ by the definition of } t^*(q_1)\\ &<(B(t)-(1-f(t)))\mu_2(\mu_2-q_2\lambda(1-t)) \text{ by inequality (B.38)}\\ \Rightarrow B(t)\mu_2(\mu_2-q_2\lambda(1-t))&>\mu_2(\mu_2-q_2\lambda(1-t))(1-f(t))-q_2\lambda^2(1-t+F(t))((1-t)f(t)+F(t)).\end{aligned}$$

Thus at $t=t^*(q_1)$ and $q=q_2$, we have the RHS of Equation (B.37)<LHS of Equation (B.37).

(b) Next, we show that under $q=q_2$, at the point $t=\bar{t}$, we have the RHS of Equation (B.37)>LHS of Equation (B.37). This is because at $q=\tilde{q}$ and $t=\bar{t}$, by definition of $\tilde{q}$, we have the RHS of Equation (B.37)=LHS of Equation (B.37) because Equation (B.31) and Equation (B.37) are equivalent. Because $q_2<\tilde{q}$, by rearranging Equation (B.31) at $t=\bar{t}$, we get the RHS of Equation (B.37)>LHS of Equation (B.37) at $q=q_2$.

Combining the results in Part (a) and Part (b), it must be true that there exists $t^*(q_2)\in(\bar{t},t^*(q_1))$ that solves Equation (B.37). Thus the induction result holds. Q.E.D.

B.10. Proof of Proposition 7

Proof: We first state a Corollary that has been proved within the proofs of Lemma B.8 and Lemma B.9 above.

Corollary B.1 *When $t^* < \bar{t}$, the solution $t(q)$ to Equation* (B.35) *always satisfies $t(q) \geq t^*$ for all $q \in [0,1]$. When $t^* > \bar{t}$, the solution $t(q)$ to Equation* (B.37) *always satisfies $t(q) \leq t^*$ for all $q \in [0,1]$.*

With Corollary B.1, we only need to show that the solutions under the priority rule are unique among all $t(q) \geq t^*$ when $t^* < \bar{t}$, and among all $t(q) \leq t^*$ when $t^* > \bar{t}$. We start with the first case.

Case 1. When $t^* < \bar{t}$, recall (B.25) is

$$\begin{aligned}&\frac{\mu_2(\mu_2 - \lambda(1-t+F(t)))(\mu_2 - q\lambda F(t))}{\mu_1 - \lambda t} \\ &= \mu_2^2(1-(1-q)f(t)) - q(\mu_2 - \lambda(1-t+F(t)))((\mu_2 + \lambda(1-t))f(t) + \lambda F(t)).\end{aligned} \tag{B.39}$$

Define

$$A(t,q) := \frac{\mu_2(\mu_2 - \lambda(1-t+F(t)))(\mu_2 - q\lambda F(t))}{\mu_1 - \lambda t}.$$

We take the derivative of the RHS of Equation (B.39) with respect to t, and get

$$-\mu_2^2(1-q)f(t) - q\lambda(1-f(t))((\mu_2+\lambda(1-t))f(t)+\lambda F(t)) - q(\mu_2 - \lambda(1-t+F(t)))(\mu_2+\lambda(1-t))f'(t) < 0.$$

Hence, the RHS is strictly decreasing in t. We next take the derivative of $A(t,q)$ with respect to t, and get

$$\frac{\lambda\mu_2((\mu_2-\lambda(1-t+F(t))+(1-f(t))(\mu_1-\lambda t))(\mu_2-q\lambda F(t)) - q(\mu_2-\lambda(1-t+F(t)))(\mu_1-\lambda t)f(t))}{(\mu_1-\lambda t)^2}. \tag{B.40}$$

If (B.40) is non-negative, which implies that the LHS is increasing in t, we can conclude that only one solution exists. (B.40) ≥ 0 is equivalent to

$$q \leq \frac{\mu_2(\mu_2-\lambda(1-t+F(t))+(1-f(t))(\mu_1-\lambda t))}{\lambda F(t)(\mu_2-\lambda(1-t+F(t))+(1-f(t))(\mu_1-\lambda t))+f(t)(\mu_2-\lambda(1-t+F(t)))(\mu_1-\lambda t)} := Y(t).$$

We need to show that $Y(t) \geq q$. We first show $Y(t)$ is decreasing with respect to t. The numerator of $Y'(t)$ is

$$\begin{aligned}-\mu_2(\mu_1-\lambda t)f'(t)f(t)(\mu_2-\lambda(1-t+F(t)))(\mu_1-\lambda t) - \Big(\mu_2(\mu_2-\lambda(1-t+F(t))+(1-f(t))(\mu_1-\lambda t))\Big)& \\ \times\Big(\lambda f(t)(1-f(t))(\mu_1-\lambda t)+f'(t)(\mu_2-\lambda(1-t+F(t)))(\mu_1-\lambda t)+f(t)\lambda(1-f(t))(\mu_1-\lambda t)\Big) < 0,&\end{aligned}$$

which is strictly negative because all terms within the bracket are positive. Thus, $Y(t)$ is decreasing in t.

Next, we derive the conditions under which $Y(t) \geq 1$ so that $Y(t) \geq q$ for all $q \in [0,1]$. To this end, we discuss the lower bound of $\inf Y(t)$. Notice that when $\mu_1 \leq \lambda$, $\sup\{t : 0 \leq t \leq 1, \mu_1 - \lambda t > 0\} = \mu_1/\lambda$. Then $\inf Y(t) = \mu_2/(\lambda F(\mu_1/\lambda)) \geq 1$, because $\mu_2 - \lambda(1-\mu_1/\lambda + F(\mu_1/\lambda)) \geq 0 \Rightarrow \mu_2 \geq (\lambda F(\mu_1/\lambda))$. Otherwise, when $\mu_1 > \lambda$, $\max\{t : 0 \leq t \leq 1, \mu_1 - \lambda t > 0\} = 1$. At $t = 1$, $Y(1) = \mu_2/(\mu_1 - (1-\mathbb{E}[X])\lambda)$. Thus, if $\mu_2/(\mu_1 - (1 - \mathbb{E}[X])\lambda) \geq 1 \Rightarrow \mu_2 \geq \mu_1 - (1-\mathbb{E}[X])\lambda$, we have $\inf Y(t) \geq 1$.

To conclude, two cases can occur.

1.1 When $\mu_1 \leq \lambda$ or $\mu_2 \geq \mu_1 - (1-\mathbb{E}[X])\lambda$, $A(t,q)$, the LHS of Equation (B.39) is increasing in t within the stable region. Because the RHS of Equation (B.39) strictly decreases in t, the solution $t^*(q)$ to Equation

(B.39) is unique for all $q \in [0,1]$.

1.2 Otherwise, when $\mu_1 > \lambda$, i.e., $\phi > 1$, and $\mu_2 < \mu_1 - (1-\mathbb{E}[X])\lambda$, i.e., $\beta < \phi - (1-\mathbb{E}[X])$, we show Equation (B.27) has a unique solution among all $t \geq t^*$ if the following condition

$$q < \frac{\mu_2^2(\mu_2 - \lambda\mathbb{E}[X])}{\lambda\mathbb{E}[X](\mu_2(\mu_2 - \lambda\mathbb{E}[X]) + \lambda\mathbb{E}[X](\mu_1 - \lambda))} := \bar{q}, \tag{B.41}$$

is satisfied. Condition (B.41) is equivalent to $\mathbb{E}[X]/(\beta - \mathbb{E}[X]) < \beta(\alpha\beta - \mathbb{E}[X])/(\mathbb{E}[X](\phi - 1))$.

First, we rewrite Equation (B.27) as

$$\frac{(1-t)f(t) + F(t)}{\beta - (1-t+F(t)) - (1-f(t))(\phi - t)} = \frac{\beta(\alpha\beta - F(t))}{(1-t+F(t))(\phi - t)}, \tag{B.42}$$

and show Equation (B.42) has a unique solution among all $t \geq t^*$. We start by analyzing the LHS of Equation (B.42). The derivative of the LHS of Equation (B.42) equals

$$\frac{(1-t)f'(t)(\beta - (1-t+F(t)) - (1-f(t))(\phi - t))) - ((1-t)f(t) + F(t))(2(1-f(t) + (\phi - t)f'(t)))}{(\beta - (1-t+F(t)) - (1-f(t))(\phi - t))^2} \tag{B.43}$$

When $\phi > 1$,

$$\text{(B.43)} < (\phi - t)\frac{\beta - (1-t+F(t)) - (1-f(t))(\phi - t) - (1-t)f(t) - F(t)}{(\beta - (1-t+F(t)) - (1-f(t))(\phi - t))^2}f'(t) \tag{B.44}$$

$$-\frac{2((1-t)f(t) + F(t))(1-f(t))}{(\beta - (1-t+F(t)) - (1-f(t))(\phi - t))^2}, \tag{B.45}$$

where term (B.45) is negative and term (B.44) is negative when

$$\mathcal{L}(t) := \beta - (1-t+F(t)) - (1-f(t))(\phi - t) - (1-t)f(t) - F(t) < 0.$$

In addition, we have

$$\mathcal{L}'(t) = 2(1-f(t)) + (\phi - 1)f'(t) > 0.$$

$\mathcal{L}(t)$ is increasing in t and at $t = 1$, it reduces to $\beta - 2\mathbb{E}[X]$. If $\beta \leq 2\mathbb{E}[X]$, the derivative of the LHS of Equation (B.42) is always less than 0, thus it is decreasing in t. Otherwise, because at $t = t^*$, $\mathcal{L}(t^*) = -(1-t^*)f(t^*) - F(t^*) < 0$ by Equation (B.7), and there exists t' that solves $\mathcal{L}(t) = 0$. We have $\mathcal{L}(t) \leq 0$ when $t \leq t'$, and $\mathcal{L}(t) > 0$ otherwise. To conclude, if $t \leq t'$, the LHS of Equation (B.42) is strictly decreasing with respect to t. However, when $t > t'$ the LHS of Equation (B.42) might increase in t.

Next, we show that the LHS of Equation (B.42) is below a strictly decreasing function $\mathcal{R}(t)$ (note that the LHS is not always monotonic, yet we still bound it by the function below). To see this,

$$\frac{(1-t)f(t) + F(t)}{\beta - (1-t+F(t)) - (1-f(t))(\phi - t)} \leq \frac{1-t+F(t)}{\beta - (1-t+F(t)) - (1-f(t))(\phi - t)} = \frac{1}{-1 + \frac{\beta - (1-f(t))(\phi - t)}{1-t+F(t)}} := \mathcal{R}(t),$$

where the first inequality follows from $(1-t)f(t) \leq 1-t$. Note $\mathcal{R}(t)$ is strictly positive because the LHS is positive. So its denominator is positive. The denominator is continuous and strictly increasing in t because $(\beta - (1-f(t))(\phi - t))' > 0$ and $(1-t+F(t))' < 0$. By the property of the power function with the power equals -1, the original function is strictly decreasing and convex in t. So, the LHS of Equation (B.42) is always below a strictly decreasing and convex function. Thus, to prove a unique solution of Equation (B.42), it is sufficient to prove the following equation has only one solution.

$$\frac{1-t+F(t)}{\beta - (1-t+F(t)) - (1-f(t))(\phi - t)} = \frac{\beta(\alpha\beta - F(t))}{(1-t+F(t)))(\phi - t)}. \tag{B.46}$$

The value of the LHS is ∞ at $t = t^*$ because the denominator is 0 by (B.7), while the value of the RHS at $t = t^*$ is finite, so that at $t = t^*$, we have LHS>RHS. At $t = 1$, the LHS= $\mathbb{E}[X]/(\beta - \mathbb{E}[X])$, and the RHS= $\beta(\alpha\beta - \mathbb{E}[X])/(\mathbb{E}[X](\phi - 1))$, thus we have LHS<RHS under the condition

$$\frac{\mathbb{E}[X]}{\beta - \mathbb{E}[X]} < \frac{\beta(\alpha\beta - \mathbb{E}[X])}{\mathbb{E}[X](\phi - 1)}.$$

Therefore, with the condition above, there is at least one solution to (B.46).

The derivative of the RHS of Equation (B.46) equals

$$\frac{-\beta f(t)(1-t+F(t))(\phi-t)+\beta(\alpha\beta-F(t))((1-f(t))(\phi-t)+1-t+F(t))}{(1-t+F(t))^2(\phi-t)^2}, \tag{B.47}$$

the derivative of the numerator of term (B.47) equals

$$\begin{aligned}
&-\beta f'(t)(1-t+F(t))(\phi-t)+\beta f(t)(1-f(t))(\phi-t)\\
&+\beta f(t)(1-t+F(t))-\beta f(t)((1-f(t))(\phi-t)+1-t+F(t))-\beta(\alpha\beta-F(t))(2(1-f(t))+(\phi-t)f'(t))\\
=&-\beta f'(t)(1-t+F(t))(\phi-t)-\beta(\alpha\beta-F(t))(2(1-f(t))+(\phi-t)f'(t))<0.
\end{aligned}$$

When $t \to 0$, the numerator converges to $\alpha\beta^2(\phi+1) > 0$, and when $t \to 1$, the numerator converges to $\beta\mathbb{E}[X](\alpha\beta - \mathbb{E}[X] - (\phi - 1))$. Thus, if $\alpha\beta - \mathbb{E}[X] \geq \phi - 1$, the numerator of term (B.47) is always positive, and the RHS of Equation (B.46) is strictly increasing with respect to t. Otherwise, there exists t_0 that solves (B.47)$=0$, i.e., t_0 is the solution to

$$f(t)(1-t+F(t))(\phi-t) = (\alpha\beta - F(t))((1-f(t))(\phi-t)+1-t+F(t)).$$

Because the numerator of term (B.47) is strictly decreasing, it is strictly greater than 0 when $t < t_0$ and it is strictly less than 0 otherwise. Thus, the RHS of Equation (B.46) is strictly increasing with respect to t when $t < t_0$, and strictly decreasing with respect to t when $t > t_0$.

When RHS of Equation (B.46) is strictly increasing with respect to t, e.g., when $\beta - \mathbb{E}[X] \geq \phi - 1$, there is only one unique solution to Equation (B.46). Otherwise, we show that when the RHS of Equation (B.46) is strictly decreasing in t, it is concave. Let $x(t) := \beta(\alpha\beta - F(t)) > 0$, and $y(t) := (1-t+F(t))(\phi-t) > 0$. It is easy to check $x'(t) < 0$, $x''(t) < 0$, $y'(t) < 0$, and $y''(t) > 0$. To simplify the notations, we omit the variable t. The second derivative of the RHS is

$$\left(\frac{x}{y}\right)'' = \frac{(x''y - xy'')y^2 - (x'y - xy')2yy'}{y^4} < 0,$$

because $x''y - xy'' < 0$ and $x'y - xy' < 0$ when the RHS is strictly decreasing. Recall the LHS of Equation (B.46) is strictly decreasing and convex. To make sure the solution to Equation (B.46) is not obtained where $t > t_0$, we only need to ensure at $t = 1$, we have LHS<RHS. That is

$$\frac{\mathbb{E}[X]}{\beta - \mathbb{E}[X]} < \frac{\beta(\alpha\beta - \mathbb{E}[X])}{\mathbb{E}[X](\phi - 1)},$$

or equivalently

$$q < \frac{\beta^2(\beta - \mathbb{E}[X])}{\beta\mathbb{E}[X](\beta - \mathbb{E}[X]) + \mathbb{E}[X]^2(\phi - 1)} = \frac{\mu_2^2(\mu_2 - \lambda\mathbb{E}[X])}{\lambda(\mu_2\mathbb{E}[X](\mu_2 - \lambda\mathbb{E}[X]) + \lambda\mathbb{E}[X]^2(\mu_1 - \lambda))} = \bar{q},$$

which is the condition stated in case 1(b) in Proposition 7.

Case 2. When $t^* > \bar{t}$, we define $\hat{q}(t)$ as

$$\hat{q}(t) := \frac{\mu_2((1-t+F(t))(1-t)f'(t) - ((1-t)f(t)+F(t))(1-f(t)))}{\lambda(((1-t)f(t)+F(t))^2 + (1-t+F(t))(1-t)^2 f'(t))}.$$

The proof follows from Equation (B.37). Recall the LHS of Equation (B.37), i.e., $B(t)$, is strictly increasing in t. The RHS of Equation (B.37) is the sum of two parts: 1) $1-f(t)$, which is strictly decreasing in t, and 2) $Z(t)$, which is defined as

$$Z(t) := -\frac{q\lambda^2(1-t+F(t))((1-t)f(t)+F(t))}{\mu_2(\mu_2 - q\lambda(1-t))}.$$

We show by the Lemma B.10 below, as long as $q < \hat{q}(t)$ for all $t < t^*$, $Z(t)$ is decreasing with respect to t within the interval $[0, t^*)$. Thus the RHS of Equation (B.37), which is the sum of two decreasing functions, is also decreasing in t. Therefore, there is at most one solution to Equation (B.37), i.e., the equilibrium is unique.

Lemma B.10 *If $q < \hat{q}(t)$, $Z'(t) < 0$.*

The proof of Lemma (B.10) is provided in Section B.10.1. Q.E.D.

B.10.1. Auxiliary Results Used in the Proofs of Proposition 7

Proof of Lemma B.10:

$$\begin{aligned} Z'(t) = & -\frac{q\lambda^2(-(1-f(t))((1-t)f(t)+F(t)) + (1-t+F(t))(1-t)f'(t))\mu_2(\mu_2 - q\lambda(1-t))}{\mu_2^2(\mu_2 - q\lambda(1-t))^2} \\ & + \frac{q\lambda^2(1-t+F(t))((1-t)f(t)+F(t))\mu_2 q\lambda}{\mu_2^2(\mu_2 - q\lambda(1-t))^2} \end{aligned}$$

The numerator of $Z'(t)$ equals

$$\mu_2 q\lambda^2\Big(((1-f(t))(\mu_2 - q\lambda(1-t)) + q\lambda(1-t+F(t)))((1-t)f(t)+F(t)) - (1-t+F(t))(1-t)f'(t)(\mu_2 - q\lambda(1-t))\Big).$$

Rearranging the term within the big bracket and collecting q, we get

$$q\lambda\Big(((1-t)f(t)+F(t))^2 + (1-t+F(t))(1-t)^2 f'(t)\Big) - \mu_2\Big((1-t+F(t))(1-t)f'(t) - ((1-t)f(t)+F(t))(1-f(t))\Big).$$

Under the condition

$$q < \frac{\mu_2\Big((1-t+F(t))(1-t)f'(t) - ((1-t)f(t)+F(t))(1-f(t))\Big)}{\lambda\Big(((1-t)f(t)+F(t))^2 + (1-t+F(t))(1-t)^2 f'(t)\Big)} = \hat{q}(t),$$

we have $Z'(t) < 0$. Q.E.D.