RL or URL: Managing Outpatient (Tele)visits with Strategic Behavior

Nan Liu
Carroll School of Management, Boston College, Chestnut Hill, MA 02467, USA, nan.liu@bc.edu

Shan Wang
School of Business, Sun Yat-sen University, Guangzhou, 510275, China, wangsh337@mail.sysu.edu.cn

Noa Zychlinski
Faculty of Data and Decision Sciences, Technion – Israel Institute of Technology, Haifa 3200003, Israel noazy@technion.ac.il

Problem definition: Many outpatient care providers offer virtual service that patients can access via televisits. Televisits allow patients to wait to be seen in the location of their choice, protected from exposure to ill patients and without going through the trouble of physical travel. There is evidence, however, that televisits are more likely to lead to a supplementary in-person visit, consuming additional resources that could have been saved if the patient's initial visit was in-person. Given this trade-off, we study whether an outpatient care provider should adopt virtual service and, if so, how best to manage a practice that simultaneously offers both virtual (or URL) and in-person (real-life, or RL) services. **Methodology/results:** We develop a stylized queueing-game model, which incorporates patient strategic choice between the two service channels. We study how a revenue-driven provider should allocate capacity between these two channels and how to incentivize patients for in-person visits. We find that the size of the system, measured by the total available service capacity relative to total patient demand, plays a determining role here. Small and large systems are better off focusing on one channel only and have no need to use in-person incentives, whereas medium-sized systems can benefit from offering both channels and in-person incentives. We also find that overall patient access to care may be hurt with the use of in-person incentives, unless the payment gap between the two channels is significantly large. **Managerial implications:** Despite the growing adoption of telemedicine, offering virtual service may not be the best choice for all providers. Capacity coordination between the virtual and in-person service channels has to be carefully balanced. Furthermore, in-person incentives need to be used with caution, otherwise patient access to care may be impaired. Proper financial incentives set up by the payer may prevent such a negative outcome.

Key words: healthcare operations management, telemedicine, queueing game, capacity management

1. Introduction

The COVID-19 pandemic accelerated the use of virtual services in general and telemedicine in the context of healthcare services in particular (Bokolo 2020, Kadir 2020). Telemedicine enables the provision of remote clinical services via real-time communication between patients and healthcare providers through video conferencing and remote monitoring (Monaghesh and Hajizadeh 2020). Virtual services, which save traveling costs and are associated with lower waiting costs for patients, can increase the efficiency of healthcare delivery (Wong et al. 2021, Kadir 2020, Hur and Chang 2020). These virtual services, however, may also have an inherent potential for providing low-value instead of quality care (O'Reilly-Jacob et al. 2021). Indeed, research has shown that a virtual visit may lead to a supplementary in-person visit for the same medical concern within a short period of time due to misdiagnosis and/or inadequate treatment (McConnochie et al. 2015, Ashwood et al. 2017, Shi et al. 2018). This situation can have a negative impact on a provider's operational efficiency. For instance, Li et al. (2021) found that when using telemedicine, on-demand virtual care increases follow-up care and, therefore, episode costs. By analyzing a three-year data set from a large payer, they found that patients with initial visits for acute respiratory infection were more likely to seek follow-up care within seven days after a virtual visit than after in-person visits (10.3% vs. 5.9%). In addition, Bavafa et al. (2018) found that e-visits trigger about 6% more office visits and that physicians accept 15% fewer new patients each month following the adoption of e-visits. These results, however, are not always consistent; for example, Reed et al. (2021) found that e-visits lead to only slightly more supplementary in-person visits compared with an initial conventional in-person visit.

Despite the mixed results on health outcomes, telemedicine will likely continue to grow in the wake of COVID-19. It can never, however, replace in-person visits (Rosenthal 2021). In-person care can offer benefits beyond virtual care, such as certain diagnosis/tests that can only be done in person, better communication, and a closer doctor-patient relationship. During the height of the pandemic, there was payment parity between telehealth and in-person care. That is, televisits were covered at the same rate as if they were in-person visits. With the end of the pandemic in sight, in-person visits are regaining popularity from patients as well as providers. This shift has led to ongoing policy discussion that focuses on payment equity rather than parity in the post-pandemic era (Shachar et al. 2020). In other words, the general opinion is that reimbursing virtual visits, which tend to be shorter and include fewer diagnostic services than in-person visits, at identical rates as in-person visits represents over-payment. In the US, these higher reimbursement rates are scheduled to end soon unless lawmakers choose to extend the payment parity policy.

Facing these changes, providers have started re-orienting their resources toward in-person visits. A 2022 McKinsey report reveals that as the pandemic abates, more physicians are gravitating away from virutal care and would prefer a return to in-person care (Cordine et al. 2022). There have been a 13% increase in physicians recommending in-person visits over telehealth and 10% increase in physicians offering in-person care only since July 2020. Providers are endeavoring to encourage in-person visits, and they take efforts in a variety of forms such as improving the in-clinic encounter experience, investing in a better physical service environment, and issuing a transportation bonus (e.g., a parking voucher, a day pass for subways, and compensation for transportation costs; the reader is referred to entire issue focusing on offering (free) transportation services using a third-party service (The American Academy of Family Physicians 2019)). For ease of discussion, we use incentive (or bonus) as an umbrella term for such efforts.

Our research is motivated by the changing landscape of telemedicine and seeks to understand how an outpatient care provider can best run a practice that offers both in-person (real-life, RL) and virtual (or URL) services. We focus on the setting of an urgent care center or a community clinic that patients visit only when they feel sick. An example of such a setting, mentioned in Cakici and Mills (2022), is Northwell Health, the largest healthcare provider in New York State. When patients seek urgent care, they are offered two options – coming in for an in-person visit or booking a virtual visit. Patients weigh these two alternatives before deciding on one. The utilities/costs associated with these two options are different. In-person visits require waiting in the clinic, whereas patients, who opt for the virtual visit, have the option to stay in the comfort of their home or any alternative location the patient prefers. Hence the waiting cost rate for virtual visits is smaller. Moreover, there are traveling costs (e.g., transportation and parking costs as well as time) associated with in-person visits. A virtual visit, however, comes with the catch of a possible supplementary in-person visit that requires the patient to travel to the clinic, join the queue of patients waiting for an in-person visit, and then wait for the next available service slot.

To efficiently manage these two channels for outpatient care, providers have two operational levers. The first is to allocate capacity; that is, given the daily service capacity, set the number of service slots for each channel. Providing more service slots will decrease the average waiting time in that channel, and consequently, influence patients' perceived utilities regarding both channels. The second operational lever is to incentivize patients for their (first and second) in-person visits, as discussed above.

We study how providers should use these two operational levers. In particular, we develop a stylized queueing-game model to determine the optimal capacity allocation and the use of in-person visit incentives in an urgent care center, where patients can choose between the in-person channel or virtual channel. In our model, the provider faces an exogenous stream of strategic patient demand. A strategic patient considers the anticipated waiting in each channel and the probability of needing a supplementary in-person visit after a virtual one. The patient has three options – walking-in for an in-person visit, booking a virtual visit, and balking. If an in-person visit is chosen, the patient's utility is the service reward net the waiting cost and traveling cost associated with the in-person channel. If a virtual visit is chosen, the patient's utility incorporates the waiting cost in the virtual channel and the expected utility of needing a possible supplementary in-person visit, which will incur additional waiting and traveling costs to attend the in-person visit. Lastly, if the patient balks, the utility is normalized to zero. In the model, patient choice is endogenous to the provider's capacity allocation because the waiting cost is influenced by the service capacity in each channel. The provider seeks to maximize the total revenue from both channels, by judiciously allocating capacity to each channel (and setting up incentives for in-person visits).

We first show that given any capacity allocation, there exists a unique mixed-strategy equilibrium in patients' choice. Since patients have three alternatives (televisit, in-person visit, or balk), there are, in total, seven equilibrium regions depending on how patients mix their choices. With the goal of maximizing revenue, we fully characterize the optimal allocation of capacity between the two service channels, where the in-person channel serves both first in-person visits and the second (supplementary) visits of patients from the virtual channel. The optimization incorporates patient equilibrium, which is endogenous to the capacity allocation. We find that the optimal system configuration (i.e., which channels to utilize and how much capacity to allocate to each) depends on the size of the system measured by the total available service capacity. In particular, when the exogenous stream of patient demand is fixed, a provider who has either limited or abundant service capacity (relative to patient demand) achieves a higher revenue when focusing on only one channel. When focusing on the virtual channel, some capacity must still be reserved for the in-person channel to serve patients who require a supplementary visit following a virtual one. Medium-sized systems, in contrast, can achieve higher revenue when providing both service channels and allowing patients to strategically choose their preferred channel.

To further improve operational performance, the provider can set up incentives for in-person visits to attract more patients. We fully characterize the joint optimal decision for capacity allocation and the use of incentives. It is natural to see that the use of in-person incentives increases total revenue. What is more interesting is that adopting incentives can fundamentally change the optimal system design. For example, a system optimally designed to focus on the virtual channel without incentives may be optimized by changing to the in-person channel (and vice versa) with an in-person visit incentive. One would expect the use of incentives to increase the total number of patients served (i.e., the total rate at which patients access these two service channels, or simply put, the total access rate) because the incentive is set to attract more in-person visits, which do not require supplementary visits and hence do not need to draw additional service resources. This, however, is not always true; the use of in-person incentives may actually lead to a decrease in the total access rate. The driving force behind this "backfire" is the potential shift of the optimal system design under in-person incentives: the provider may opt to serve fewer patients who come with a higher payment rate. To avoid such a negative impact on social welfare, one solution we identify is in line with the idea of "payment equity" discussed above. We show that if the payment rates for these two types of visits are sufficiently differentiated, the total access rate will not decrease with the adoption of in-person incentives by a revenue-driven provider.

The rest of the paper is organized as follows. In Section 2 we briefly review the relevant literature. In Section 3, we present the basic model and introduce the provider's objective in regard to capacity allocation decisions for the two service channels. Section 4 analyzes the different equilibrium regions and the optimal capacity allocation. In Section 5, we study the option of providing incentives for in-person visits. Section 6 presents concluding remarks and a few relevant directions for future research. All proofs of the technical results are shown in the Online Appendix.

2. Literature Review

This research draws upon the literature on healthcare operations management (OM) and queueing studies with strategic customers. We review each stream below.

2.1. Healthcare OM

Within the healthcare OM literature, our work is related to the research that (1) uses queueing models to investigate system design questions in outpatient care and (2) studies multi-channel care settings, in particular where one channel offers the telemedicine option. Two of the studies from the first stream of literature include Green and Savin (2008), Zacharias and Armony (2016). They developed stylized queueing models to investigate system design questions such as panel size selection and capacity decisions. An interesting phenomenon addressed by this literature is that patients may need to revisit the provider after (no-shows at) initial visits. Although this patient "returning" feature also presents in our model, our patient revisits require a different source of capacity compared to their initial visits. In addition, patients choose between two channels to access care in our model, while the prior literature mostly focuses on a single channel for in-person visits.

Recently, a few papers addressed OM issues related to multiple access channels for outpatient care (e.g., in-person appointment visits, walk-ins, virtual visits, visits to different providers). An outpatient care setting where strategic patients choose between a walk-in visit and an appointment was studied in Liu et al. (2023). They characterized the different equilibrium cases and allocated capacity to the two channels. In their paper, only walk-in patients may balk. Huang et al. (2022) studied the doctor-shopping behavior (i.e., patients seek opinions from multiple doctors without referrals) and its impact on social welfare. In the present work, we consider different levels of quality of care between the channels, captured by the need for a supplementary in-person visit following a virtual visit.

As telemedicine and video consultation gain popularity, research studying their effect and patients' preferences has started emerging. A telehealth setting for chronic strategic patients was studied by Rajan et al. (2019). They found that telehealth increases both the access rate and provider revenue. This is not necessarily true, however, when considering supplementary in-person visits. In particular, as we and Cakici and Mills (2022) show, maximizing provider revenue may hurt patient access when considering in-person supplementary visits. Bavafa et al. (2021) studied a chronic-care setting, incorporating telehealth and office visits. Their goal was to set the time interval between consecutive visits while investigating the impact of different reimbursement schemes on patient panel size, physician earnings, and overall patient health. They showed that e-visits may have a negative effect in terms of panel size/health.

Recently, Zychlinski (2023) and Cakici and Mills (2022) studied a similar setting to ours with in-person virtual and returning channels. Using a fluid model approximation, Zychlinski (2023) addressed scheduling and capacity allocation decisions among the three service channels: in-person, virtual and supplementary service for returning virtual patients. They developed an index-based policy that demonstrates the importance of capacity coordination among the channels. Using a three-stage game theoretic model, Cakici and Mills (2022) studied the effect of telehealth reimbursement policies on patient access to acute care. In contrast to Cakici and Mills (2022), our focus here is on the operational aspects of managing such hybrid systems rather than on the financial/reimbursement aspect. In particular, we fully characterize the optimal system design in terms of capacity allocation for different system sizes. Then, we study the effect of utilizing an in-person incentive on revenue and total access rate. In a sense, Cakici and Mills (2022) and our work complement each other and contribute to the understanding of hybrid healthcare settings.

2.2. Queueing Studies with Strategic Customers

In terms of methodology, our research is based on the analysis of queuing systems with strategic customers. This broad research area, which started with Naor (1969), considers customer join/balk decisions based on how sensitive customers are to waiting as well as optimizing system efficiency/social welfare by determining the service capacity, pricing, or priority schemes. Hassin and Haviv (2003) offered a comprehensive review of this area. Below we draw attention to studies that are most relevant to our work.

Hassin and Roet-Green (2020) studied a service system that requires customers to travel to the queue to be served. Specifically, customers observe the queue length and then decide whether to travel. In our work, we consider a virtual service channel that requires no traveling and an in-person channel (first- and second-time visitors) that requires traveling. Furthermore, we optimize the travel cost by considering a (transportation) bonus to increase revenue. Another recent relevant paper that considered a traveling cost is that by Baron et al. (2022) who studied an omnichannel service system that includes walk-in and online service channels. They showed that although the online channel increases revenue, in equilibrium it also reduces customers' individual utility and social welfare. To overcome this, the authors suggested prioritizing walk-in customers; this benefits the service provider and customers in equilibrium compared with only having a walk-in service channel.

There are different ways to improve customer service experience. Baron et al. (2014), for example, used strategic idling to study scheduling policies, mainly in healthcare systems where patients have to go through several diagnostic and treatment stations. We suggest and analyze a different approach: incentives to attract in-person patients that compensate them for their traveling costs/burden. We show that while using these incentives is likely to increase revenue, it might also harm patients' service access levels.

Finally, we conclude this section by summarizing our contributions. We develop the modeling framework for an outpatient care provider who offers both in-person and virtual services. Our model captures the key trade-off strategic patients have to make when choosing between these two channels for service.

An important and interesting feature of our model is that it operationally captures the difference in care quality of these two channels: patients who visit the virtual channel may need a second supplementary in-person visit. We first prove the existence and uniqueness of a mixed patient strategy in equilibrium, based on which we characterize the optimal capacity allocation and use of in-person incentives for a revenue-driven provider. We find that the size of the system, measured by the total available service capacity relative to total patient demand, plays a critical role in determining the provider's optimal decisions. Small and large systems are better off focusing on one channel only and have no need to use in-person incentives, whereas medium-sized systems can benefit from offering both channels and in-person incentives. We advise caution when using in-person incentives, which, as we demonstrate, may backfire and hurt overall patient access to care.

3. The Basic Model

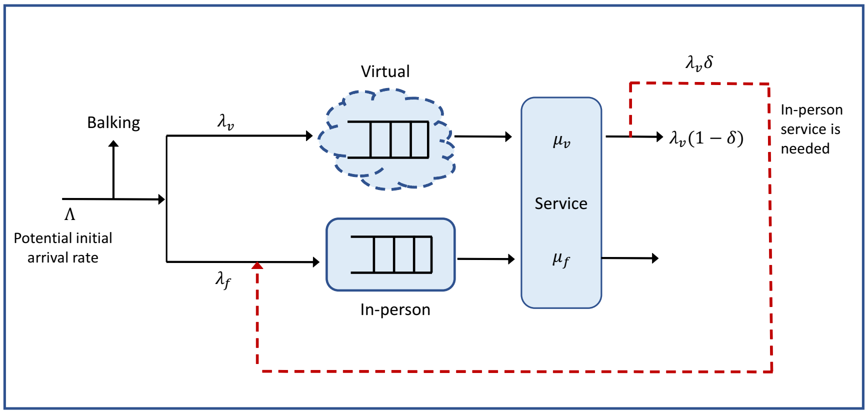
A provider offers two service channels to their patients: in-person (i.e., face-to-face) and virtual. The provider's daily service capacity is fixed at μ service slots (all slots have equal length, e.g., 10 minutes). They need to decide how many slots to allocate to the face-to-face channel (denoted by μ_f) and the virtual channel (denoted by μ_v) such that $\mu_f + \mu_v = \mu$. Note that the actual service time of patients may have some variability, but the provider is expected to be able to serve μ_f in-person patients and μ_v virtual patients per day.

Homogeneous strategic patients arrive following a Poisson process with daily rate Λ. Offered these two channels, each patient has three options: using the face-to-face channel, the virtual channel, or balking. As patients are homogeneous, they will use the same mixed strategy in the equilibrium, i.e., choosing the in-person channel with probability p_f, choosing the virtual channel with probability p_v and balking with probability p_b if such a equilibrium exists. Let $\lambda_f = p_f\Lambda$ and $\lambda_v = p_v\Lambda$ denote the corresponding arrival rates for the in-person and virtual channels, respectively. Since the virtual channel does not have the same effectiveness as the face-to-face channel, some patients need a supplementary in-person service visit with the same provider after their virtual visit. The supplementary visit happens with probability δ, which reflects the quality of the virtual service (a smaller δ indicates better service quality). To avoid trivialities, we assume that $\delta \in (0, 1)$. We further assume that face-to-face visits always resolve patient health issues.

To analyze patient strategy, we start by studying the utility of each choice. We begin with the arrival–service process, which is modeled as a queueing network (see Figure 1).

To make the analysis more tractable, we assume that the service process is Markovian and hence the queueing network is a Jackson network. In steady state, the Jackson network performs as two M/M/1 queues, one for the virtual channel and the second for the face-to-face channel. The arrival and service rates are λ_v and μ_v in the virtual queue and $\lambda_f + \delta\lambda_v$ and μ_f in the in-person queue.

Figure 1 **The hybrid model.**



3.1. Patient's Utility

Patients make their choices based on the expected utility of each channel.

Cost of each face-to-face visit. We consider a traveling cost T that is associated with each visit (e.g., transportation cost, parking cost, etc.). In general, T represents a fixed inconvenience cost incurred by the in-person channel. Since each queue in the Jackson network can be regarded as an M/M/1 queue, the expected waiting time for each face-to-face visit is

$$W_f(\lambda_f, \lambda_v) = [\mu_f - (\lambda_f + \delta\lambda_v)]^{-1}. \tag{1}$$

Let θ_f denote the waiting cost per time unit in the face-to-face channel. Then, the expected total cost associated with each face-to-face visit is

$$C_f(\lambda_f, \lambda_v) = T + \theta_f W_f(\lambda_f, \lambda_v). \tag{2}$$

Cost of each virtual visit. The expected waiting time in the virtual channel is

$$W_v(\lambda_v) = [\mu_v - \lambda_v]^{-1}. \tag{3}$$

Let θ_v denote the waiting cost per time unit in this channel. Since patients in the virtual channel can be anywhere during their appointment, while patients in the face-to-face channel must commute to and wait in the clinic, we assume that $\theta_f \geq \theta_v$. In particular, during a pandemic outbreak, the gap between θ_v and θ_f increases, as patients in the virtual channel have a lower exposure risk to other ill patients. Though there is smaller waiting cost and no traveling cost in the virtual channel, virtual patients may need a supplementary face-to-face visit, with all the additional costs associated with it. The expected total cost of each virtual service is, therefore,

$$C_v(\lambda_f, \lambda_v) = \theta_v W_v(\lambda_v) + \delta C_f(\lambda_f, \lambda_v). \tag{4}$$

Patients receive a service reward R once their healthcare demands are satisfied. Without loss of generality, we assume that the utility of balking is 0. Then, by (1)–(4), the following utility for each alternative is:

$$\begin{aligned} U_f(\lambda_f, \lambda_v) &= R - T - \theta_f \left[\mu_f - (\lambda_f + \delta\lambda_v)\right]^{-1}, & (5)\\ U_v(\lambda_f, \lambda_v) &= R - \theta_v \left[\mu_v - \lambda_v\right]^{-1} - \delta T - \delta\theta_f \left[\mu_f - (\lambda_f + \delta\lambda_v)\right]^{-1}, & (6)\\ U_b(\lambda_f, \lambda_v) &= 0. \end{aligned}$$

Homogeneous patients will choose (λ_f, λ_v) to maximize their expected utility $(\lambda_f U_f(\lambda_f, \lambda_v) + \lambda_v U_v(\lambda_f, \lambda_v))/\Lambda$, for which the denominator can be omitted from the analysis because it is a constant.

3.2. The Provider's Problem

The provider receives a reward r_f for each patient who chooses the face-to-face channel, and a reward r_v for each patient who chooses the virtual channel. The provider's problem is to allocate capacity in anticipation of patient decisions to maximize their total compensation:

$$\begin{aligned} \max_{\mu_f, \mu_v} \quad & r_f\lambda_f + r_v\lambda_v & (7)\\ s.t. \quad & \mu_f + \mu_v = \mu & (8)\\ & \mu_f \geq 0, \mu_v \geq 0 & (9)\\ & (\lambda_f, \lambda_v) = \underset{\lambda_v + \lambda_f \leq \lambda}{\arg\max} \ \left[\lambda_f U_f(\lambda_f, \lambda_v) + \lambda_v U_v(\lambda_f, \lambda_v)\right] & (10)\\ & U_f(\lambda_f, \lambda_v),\ U_v(\lambda_f, \lambda_v) \text{ as defined in (5) and (6).} & (11) \end{aligned}$$

Remark 1. The provider's reward structure in our model is quite flexible and can capture various reimbursement regimes in healthcare. The pair (r_f, r_v) naturally represents a bundle payment mechanism because r_v covers the whole episode of care, which may involve a supplementary face-to-face visit. Under the fee-for-service payment mechanism, suppose that the payment for a virtual visit and a face-to-face one is r_v^{fee} and r_f^{fee}, respectively. The objective (7) remains valid by setting $(r_f^{\text{fee}}, r_v^{fee} + \delta r_f^{\text{fee}})$ as (r_f, r_v).

4. Patients' Equilibrium and Provider's Optimal Decision

4.1. Patients' Equilibrium Strategy

The formulation (7)-(10) is well defined if patient equilibrium exists and is unique. We establish this result in this section. We first study the effective arrival rate to each channel in equilibrium for any given capacity allocation. To illustrate the different strategies, we use B to denote the pure strategy of balking, V to denote the pure strategy of choosing the virtual channel, and F to denote the pure strategy of choosing the face-to-face channel. For mixed strategies, we use letter combinations of B, V and F. For example, we use BF to denote the mixed strategy of the face-to-face channel and balking. In total, we have seven different types of strategies: B, V, F, BVF, BF, BV and VF.

Per (10), the effective arrival rates in equilibrium (λ_f, λ_v) must satisfy the following condition for $x \in \{f, v, b\}$:

$$\text{If} \quad \lambda_x > 0, \text{then} \quad U_x(\lambda_f, \lambda_v) = \max\{U_f(\lambda_f, \lambda_v), U_v(\lambda_f, \lambda_v), U_b(\lambda_f, \lambda_v)\}.$$

That is, in equilibrium, no patient will choose a channel with lower utility than the other channel or balking. Table 1 presents the different strategies and the corresponding effective arrival rates and utilities.

Table 1 Equilibrium strategies

Strategy	Effective Arrival Rates	Utilities
B	$\lambda_f = 0, \quad \lambda_v = 0$	$U_f(0,0) \le 0, \quad U_v(0,0) \le 0$
V	$\lambda_f = 0, \quad \lambda_v = \lambda$	$U_f(0,\lambda) \le U_v(0,\lambda), \quad U_v(0,\lambda) \ge 0$
F	$\lambda_f = \lambda, \quad \lambda_v = 0$	$U_f(\lambda,0) \ge U_v(\lambda,0), \quad U_f(0,\lambda) \ge 0$
BVF	$\lambda_f + \lambda_v \le \lambda$	$U_f(\lambda_f,\lambda_v) = U_v(\lambda_f,\lambda_v) = 0$
BF	$\lambda_f \le \lambda, \quad \lambda_v = 0$	$U_f(\lambda_f,0) = 0, \quad U_v(\lambda_f,0) \le 0$
BV	$\lambda_f = 0, \quad \lambda_v \le \lambda$	$U_f(0,\lambda_v) \le 0, \quad U_v(0,\lambda_v) = 0$
VF	$\lambda_f + \lambda_v = \lambda$	$U_f(\lambda_f,\lambda_v) = U_v(\lambda_f,\lambda_v) \ge 0$

Theorem 1 establishes the existence and uniqueness of the equilibrium strategy for patients, given any service capacity allocation (μ_f, μ_v).

Theorem 1. *Consider the following cases:*

1. *When $(\mu_f, \mu_v) \in$ Region B, there exists a unique equilibrium strategy – strategy B – such that $\lambda_f = \lambda_v = 0$. Region B satisfies:*

$$\begin{aligned}
&\theta_v \mu_v^{-1} + \delta\theta_f \mu_f^{-1} \geq R - \delta T, \\
&\mu_f \leq \theta_f \left[R - T\right]^{-1}.
\end{aligned}$$

2. *When $(\mu_f, \mu_v) \in$ Region V, there exists a unique equilibrium strategy – strategy V – such that $\lambda_v = \lambda$. Region V satisfies:*

$$\begin{aligned}
&\theta_v \left[\mu_v - \lambda\right]^{-1} - (1-\delta)\theta_f \left[\mu_f - \delta\lambda\right]^{-1} \leq T - \delta T, \\
&\theta_v \left[\mu_v - \lambda\right]^{-1} + \delta\theta_f \left[\mu_f - \delta\lambda\right]^{-1} \leq R - \delta T, \\
&\mu_f \geq \delta\lambda + \delta\theta_f \left[R - \delta T\right]^{-1}, \\
&\mu_v \geq \lambda + \theta_v \left[R - \delta T\right]^{-1}.
\end{aligned}$$

3. *When $(\mu_f, \mu_v) \in$ Region F, there exists a unique equilibrium strategy – strategy F – such that $\lambda_f = \lambda$. Region F satisfies:*

$$\begin{aligned}
&\mu_f \geq \lambda + \theta_f \left[R - T\right]^{-1}, \\
&\theta_v \mu_v^{-1} - (1-\delta)\theta_f \left[\mu_f - \lambda\right]^{-1} \geq T - \delta T.
\end{aligned}$$

4. *When $(\mu_f, \mu_v) \in$ Region BVF, there exists a unique equilibrium strategy – strategy BVF – such that $\lambda_f + \lambda_v \leq \lambda$. Region BVF satisfies:*

$$\begin{aligned}
&\mu_v \geq \theta_v \left[(1-\delta)R\right]^{-1}, \\
&\mu_f - \delta\mu_v \geq \theta_f \left[R - T\right]^{-1} - \delta\theta_v \left[(1-\delta)R\right]^{-1}, \\
&\mu_f + (1-\delta)\mu_v \leq \lambda + \theta_f \left[R - T\right]^{-1} + \theta_v R^{-1}.
\end{aligned}$$

The effective arrival rates (λ_f, λ_v) can be derived from $U_f(\lambda_f, \lambda_v) = U_v(\lambda_f, \lambda_v) = 0$.

5. *When $(\mu_f, \mu_v) \in$ Region BF, there exists a unique equilibrium strategy – strategy BF – such that $\lambda_f \leq \lambda,\ \lambda_v = 0$. Region BF satisfies:*

$$\begin{aligned}
&\theta_f \left[R - T\right]^{-1} \leq \mu_f \leq \lambda + \theta_f \left[R - T\right]^{-1}, \\
&\mu_v \leq \theta_v \left[(1-\delta)R\right]^{-1}.
\end{aligned}$$

The effective arrival rate λ_f can be derived from $U_f(\lambda_f, 0) = 0$.

6. *When $(\mu_f, \mu_v) \in$ Region BV, there exists a unique equilibrium strategy – strategy BV – such that $\lambda_f = 0,\ \lambda_v \leq \lambda$. Region BV satisfies:*

$$\theta_v \mu_v^{-1} + \delta\theta_f \mu_f^{-1} \leq R - \delta T, \tag{12}$$

$$\theta_v \left[\mu_v - \lambda\right]^{-1} + \delta\theta_f \left[\mu_f - \delta\lambda\right]^{-1} \geq R - \delta T \text{ for } \mu_v > \lambda \text{ and } \mu_f > \delta\lambda, \tag{13}$$

$$\mu_f - \delta\mu_v \leq \theta_f \left[R - T\right]^{-1} - \delta\theta_v \left[(1-\delta)R\right]^{-1}. \tag{14}$$

The effective arrival rate λ_v can be derived from $U_v(0, \lambda_v) = 0$.

7. *When $(\mu_f, \mu_v) \in$ Region VF, there exists a unique equilibrium strategy – strategy VF – such that $\lambda_f + \lambda_v = \lambda$. Region VF satisfies:*

$$\theta_v \mu_v^{-1} - (1-\delta)\theta_f \left[\mu_f - \lambda\right]^{-1} \leq (1-\delta)T \quad \text{for } \mu_f > \lambda, \tag{15}$$

$$\theta_v \left[\mu_v - \lambda\right]^{-1} - (1-\delta)\theta_f \left[\mu_f - \delta\lambda\right]^{-1} \geq (1-\delta)T \quad \text{for } \mu_v > \lambda, \tag{16}$$

$$\mu_f + (1-\delta)\mu_v \geq \lambda + \theta_f \left[R - T\right]^{-1} + \theta_v R^{-1}, \tag{17}$$

$$\mu_f \geq \delta\lambda + \theta_f \left[R - T\right]^{-1}. \tag{18}$$

The effective arrival rates (λ_f, λ_v) can be derived from $\lambda_f + \lambda_v = \lambda$ and $U_f(\lambda_f, \lambda_v) = U_v(\lambda_f, \lambda_v)$.

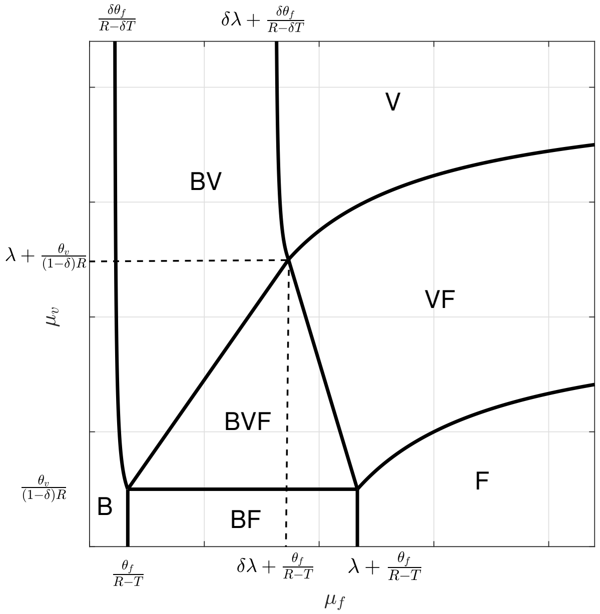
Per Theorem 1, all possible capacity allocations can be divided into seven mutually exclusive and collectively exhaustive cases. Letting the x-axis represent μ_f and the y-axis represent μ_v, Figure 2 illustrates these seven regions.

In equilibrium, when μ_f is small, no patient will choose the face-to-face channel; when μ_v is small, no patient will choose the virtual channel. When both μ_f and μ_v are small, all patients will balk. It is worth noting that μ_f has an effect on both channels' – face-to-face and virtual – utilities. In particular, when μ_f is small, those who choose virtual visits can still suffer because some of them need a supplementary face-to-face visit for which they will have to wait a long time and, accordingly, have low utility. Hence, when μ_f is small, regardless of how large μ_v is, some patients will balk. When μ_f is large, however, regardless of how small μ_v is, no patient will balk.

4.2. Optimal Capacity Allocation

Thus far, we have analyzed patients' equilibrium strategy given any capacity allocation, (μ_f, μ_v). In this section, we study the optimal capacity allocation in anticipation of patients' equilibrium. Recall the seven regions of capacity allocation. We start

Figure 2 An illustration of the seven regions and their corresponding equilibrium strategies


$\frac{\delta\theta_f}{R-\delta T}$
$\delta\lambda+\frac{\delta\theta_f}{R-\delta T}$
V
BV
$\lambda+\frac{\theta_v}{(1-\delta)R}$
VF
μ_v
BVF
$\frac{\theta_v}{(1-\delta)R}$
F
B
BF
$\frac{\theta_f}{R-T}$
$\delta\lambda+\frac{\theta_f}{R-T}$
$\lambda+\frac{\theta_f}{R-T}$
μ_f

by studying the optimal allocation in each region. Specifically, for each Region $X \in \{B, V, F, BVF, BF, BV, VF\}$, we add the constraint $(\mu_f, \mu_v) \in$ Region X to problem (7). We denote the optimal solution and the corresponding effective arrival rates subject to Region X, if any, as (μ_f^{X*}, μ_v^{X*}) and $(\lambda_f^{X*}, \lambda_v^{X*})$. Then, we compare the value function $r_f\lambda_f^{X*} + r_v\lambda_v^{X*}$ for all X to find the global optimal capacity allocation.

1. For $(\mu_f, \mu_v) \in$ Region B: All patients balk. Therefore, any feasible solution within this region, if any, is optimal and correspondingly $(\lambda_f^{B*}, \lambda_v^{B*}) = (0, 0)$.

2. For $(\mu_f, \mu_v) \in$ Region V: All patients choose the virtual channel. Therefore, any feasible solution within this region is optimal and correspondingly $(\lambda_f^{V*}, \lambda_v^{V*}) = (0, \lambda)$.

3. For $(\mu_f, \mu_v) \in$ Region F: All patients choose the face-to-face channel. Therefore, any feasible solution within this region is optimal and correspondingly $(\lambda_f^{F*}, \lambda_v^{F*}) = (\lambda, 0)$.

4. For $(\mu_f, \mu_v) \in$ Region BVF: We have $U_f(\lambda_f, \lambda_v) = U_v(\lambda_f, \lambda_v) = 0$; thus

$$\lambda_v = \mu_v - \theta_v\left[R - \delta R\right]^{-1} \quad \text{and} \quad \lambda_f + \delta\lambda_v = \mu_f - \theta_f\left[R - T\right]^{-1}.$$

Noting that the provider's total reward is linear in (λ_v, λ_f) and that in this region λ_v (λ_f) linearly increases in μ_v (μ_f) with everything else being fixed, one would expect the optimal capacity allocation to be a "bang-bang" type of control. Proposition 1 formalizes this intuition.

Proposition 1 **(optimal capacity allocation subject to Region BVF)**.

Let $[\underline{\mu}_v^{BVF}, \overline{\mu}_v^{BVF}]$ denote the range for μ_v such that $(\mu - \mu_v, \mu_v) \in$ Region BVF. The optimal capacity allocation is, accordingly, a boundary one:

- *If $r_v \geq (1+\delta) r_f$, then $(\mu_f^{BVF*}, \mu_v^{BVF*}) = (\mu - \overline{\mu}_v^{BVF}, \overline{\mu}_v^{BVF})$;*

- *Otherwise, $(\mu_f^{BVF*}, \mu_v^{BVF*}) = (\mu - \underline{\mu}_v^{BVF}, \underline{\mu}_v^{BVF})$.*

For the corresponding effective arrival rates in equilibrium, we have

$$\lambda_f^{BVF*} = \mu_f^{BVF*} - \theta_f \left[R - T\right]^{-1} - \delta \mu_v^{BVF*} + \delta \theta_v \left[R - \delta R\right]^{-1}$$

and

$$\lambda_v^{BVF*} = \mu_v^{BVF*} - \theta_v \left[R - \delta R\right]^{-1}.$$

5. For $(\mu_f, \mu_v) \in$ Region BF: we have $U_f(\lambda_f, 0) = 0$ and $\lambda_v = 0$; thus,

$$\lambda_f = \mu_f - \theta_f \left[R - T\right]^{-1}.$$

It is evident that the optimal decision is to set μ_f as large as possible. Let $\overline{\mu}_f^{BF}$ denote the upper bound for μ_f such that $(\mu_f, \mu_v) \in$ Region BF. Then,

$$(\mu_f^{BF*}, \mu_v^{BF*}) = (\overline{\mu}_f^{BF}, \mu - \overline{\mu}_f^{BF}),$$

and

$$\lambda_f^{BF*} = \overline{\mu}_f^{BF} - \theta_f \left[R - T\right]^{-1}.$$

6. For $(\mu_f, \mu_v) \in$ Region BV: we have $U_v(0, \lambda_v) = 0$ and $\lambda_f = 0$; thus,

$$R - \delta T - \theta_v \left[\mu_v - \lambda_v\right]^{-1} - \delta \theta_f \left[\mu_f - \delta \lambda_v\right]^{-1} = 0.$$

Obtaining the optimal allocation that maximizes $r_v \lambda_v$ in this region, however, is quite involved. Proposition 2 provides detailed characterizations.

Proposition 2 **(optimal capacity allocation subject to Region BV)**.

Define

$$\tilde{\mu}_v^{BV} = \left(\mu - \left(\delta(\theta_f - \theta_v) + (1-\delta)\sqrt{\delta \theta_f \theta_v}\right) \left[R - \delta T\right]^{-1}\right) \left[1 + \delta\right]^{-1}.$$

Then, we have

- *If* $\mu \geq \lambda(1+\delta) + \theta_v \left[(1-\delta)R\right]^{-1} + \theta_f \left[R-T\right]^{-1}$ *or* $(\mu - \tilde{\mu}_v^{BV}, \tilde{\mu}_v^{BV}) \in$ *Region* V*, any allocation lying on the boundary between Region* BV *and Region* V *is optimal, leading to* $\lambda_v^{BV*} = \lambda$.
- *Otherwise,*
  - *If* $(\mu - \tilde{\mu}_v^{BV}, \tilde{\mu}_v^{BV}) \in$ *Region* BV*,* $(\mu_f^{BV*}, \mu_v^{BV*}) = (\mu - \tilde{\mu}_v^{BV}, \tilde{\mu}_v^{BV})$ *and*

$$\lambda_v^{BV*} = \left(\mu - (\sqrt{\delta\theta_f} + \sqrt{\theta_v})^2 \left[R - \delta T\right]^{-1}\right) \left[1+\delta\right]^{-1};$$

  - *If* $(\mu - \tilde{\mu}_v^{BV}, \tilde{\mu}_v^{BV}) \notin$ *Region* BV*,* $(\mu_f^{BV*}, \mu_v^{BV*}) = (\mu - \underline{\mu}_v^{BV}, \underline{\mu}_v^{BV})$ *where* $\underline{\mu}_v^{BV}$ *is the lower bound for* μ_v *such that* $(\mu - \mu_v, \mu_v) \in$ *Region* BV*. Specifically,*

$$\underline{\mu}_v^{BV} = \left(\mu - \theta_f \left[R-T\right]^{-1} + \delta\theta_v \left[(1-\delta)R\right]^{-1}\right) \left[1+\delta\right]^{-1},$$

and

$$\lambda_v^{BV*} = \left(\mu - \theta_f \left[R-T\right]^{-1} - \theta_v \left[(1-\delta)R\right]^{-1}\right) \left[1+\delta\right]^{-1}.$$

7. For $(\mu_f, \mu_v) \in$ Region VF: we have $U_v(\lambda_f, \lambda_v) = U_v(\lambda_f, \lambda_v)$ and $\lambda_f + \lambda_v = \lambda$; thus,

$$\theta_v \left[\mu_v - \lambda_v\right]^{-1} - (1-\delta)\theta_f \left[\mu_f - \lambda + (1-\delta)\lambda_v\right]^{-1} = (1-\delta)T.$$

Proposition 3 characterizes the optimal allocation as well as the effective arrival rates in equilibrium subject to this region.

Proposition 3 **(optimal capacity allocation subject to Region** VF**).** *Let* $[\underline{\mu}_v^{VF}, \overline{\mu}_v^{VF}]$ *denote the range for* μ_v *such that* $(\mu - \mu_v, \mu_v) \in$ *Region* VF*.*

- ***If*** $\boldsymbol{r_v > r_f}$*,*

$$(\mu_f^{VF*}, \mu_v^{VF*}) = (\mu - \overline{\mu}_v^{VF}, \overline{\mu}_v^{VF})$$

and

$$\lambda_v^{VF*} = \min\left\{\lambda, \tfrac{1}{\delta}\left(\mu - \lambda - \theta_v \left[(1-\delta)R\right]^{-1} - \theta_f \left[R-T\right]^{-1}\right)\right\}.$$

- ***If*** $\boldsymbol{r_v = r_f}$*, any feasible allocation is optimal and the corresponding effective arrival rates in equilibrium satisfy*

$$r_f \lambda_f^{VF*} + r_v \lambda_v^{VF*} = r_f \lambda = r_v \lambda.$$

- ***If*** $\boldsymbol{r_v < r_f}$*,*

$$(\mu_f^{VF*}, \mu_v^{VF*}) = (\mu - \underline{\mu}_v^{VF}, \underline{\mu}_v^{VF}) \quad \textit{and} \quad \lambda_f^{VF*} = \lambda.$$

After deriving the optimal capacity allocation in each region, we are now ready to characterize the global optimal capacity allocation. We start by defining four system-size categories to facilitate the subsequent discussions.

Definition 1. Given the system's primitives, we consider the following system-size categories with respect to the total service capacity μ.

1. **Extremely small systems:**

$$\min\left\{\theta_f\left[R-T\right]^{-1},\underline{\mu}^{BV}\right\} \leq \mu < \max\left\{\theta_f\left[R-T\right]^{-1},\underline{\mu}^{BV}\right\},$$

where $\underline{\mu}^{BV}$ is the minimum μ required to attain Region BV, which is obtained by the solution of μ to

$$R-\delta T-\theta_v\Phi^{-1}(\mu)-\delta\theta_f\left[\mu-\Phi(\mu)\right]^{-1}=0,$$

where

$$\Phi(\mu)=\max\left\{\theta_v\left[(1-\delta)R\right]^{-1},\ \mu\left(\sqrt{\delta\theta_v\theta_f}-\theta_v\right)\left[\delta\theta_f-\theta_v\right]^{-1}\right\}. \tag{19}$$

2. **Small systems:**

$$\max\left\{\theta_f\left[R-T\right]^{-1},\underline{\mu}^{BV}\right\} \leq \mu < \lambda+\theta_f\left[R-T\right]^{-1}+\theta_v\left[(1-\delta)R\right]^{-1}.$$

3. **Medium-sized systems:**

$$\lambda+\theta_f\left[R-T\right]^{-1}+\theta_v\left[(1-\delta)R\right]^{-1} \leq \mu < (1+\delta)\lambda+\theta_v\left[(1-\delta)R\right]^{-1}+\theta_f\left[R-T\right]^{-1}.$$

4. **Large systems:**

$$\mu \geq (1+\delta)\lambda+\theta_v\left[R(1-\delta)\right]^{-1}+\theta_f\left[R-T\right]^{-1}.$$

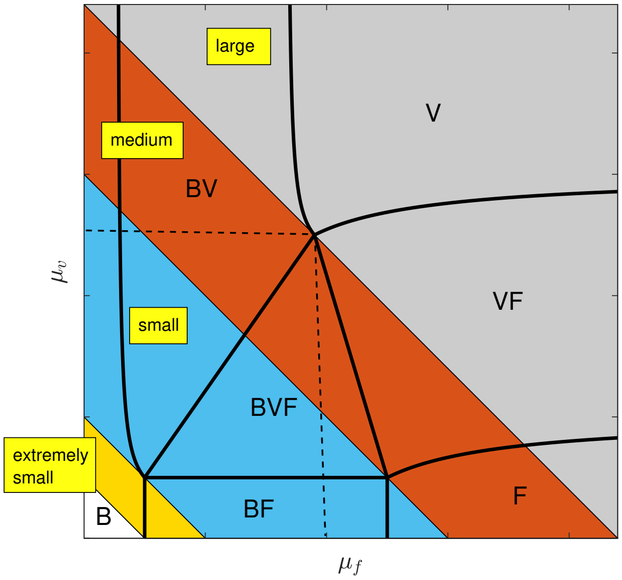
We do not consider the case where $\mu < \min\left\{\theta_f[R-T]^{-1},\underline{\mu}^{BV}\right\}$ because in this case Region B is the only feasible region, making the problem trivial. Figure 3 illustrates the four system-size categories in Definition 1.

Theorem 2 characterizes the optimal capacity allocation for each system-size category.

Theorem 2 **(optimal capacity allocation)**. *The optimal capacity allocation for each system-size category is:*

1. ***Extremely Small Systems:*** *If $\theta_f\left[R-T\right]^{-1} \leq \underline{\mu}^{BV}$, the optimal capacity allocation is in Region BF; otherwise, it is in Region BV.*

Figure 3 **An illustration of the four system-size categories.**



2. ***Small Systems:*** *If* $r_f \min\{\lambda, \mu - \theta_f [R-T]^{-1}\} \geq r_v \lambda_v^{BV*}$, *the optimal capacity allocation is in Region* BF *(note that Region* F *is a special case of Region* BF*); otherwise, it is in Region* BV*.*

3. ***Medium-sized Systems:*** *If*

$$r_v \lambda_v^{BV*} \geq r_f \lambda + \tfrac{1}{\delta}(r_v - r_f)^+ \left(\mu - \lambda - \theta_v [R(1-\delta)]^{-1} - \theta_f [R-T]^{-1}\right),$$

the optimal capacity allocation is in Region BV*; otherwise, it is in Region* VF *(note that Region* F *is a special case of Region* VF*, and dominates Region* VF *when* $r_f \geq r_v$*).*

4. ***Large Systems:*** *If* $r_v \geq r_f$*, the optimal allocation is in Region* V*; otherwise, it is in Region* F*.*

For extremely small systems, the optimal solution must be in the feasible region: either BF or BV. When $\theta_f [R-T]^{-1} \leq \underline{\mu}^{BV}$, Region BF is feasible; otherwise, Region BV is feasible.

For small systems, Region BVF is dominated by Region BV or Region BF, as the optimal solution is on the region boundary (see Proposition 1). When the face-to-face channel is more beneficial, the provider should allocate all capacity to μ_f to attract as many patients as possible to this channel. In contrast, when the virtual channel is more beneficial, the provider should attract as many patients as possible to the virtual channel by setting a large μ_v and reserving a proper amount of capacity, μ_f, for virtual patients requiring supplementary face-to-face services.

For medium-sized systems, Region BVF is sub-optimal. When r_f is large enough, attracting all patients to the face-to-face channel is more beneficial; otherwise, when r_v is large enough, attracting patients to the virtual channel (with some patients balking) is more beneficial; when the difference between r_v and r_f is not too large, mixing the two channels is more beneficial.

Lastly, for large systems, the provider should attempt to attract all patients to either virtual or face-to-face channels depending on the relation between r_v and r_f.

In summary, Theorem 2 indicates that for most cases, the provider should focus on one channel only. Serving patients from both channels is optimal only when the system is medium sized and the gap between r_v and r_f is not too large.

5. Model with In-Person Visit Incentives

Providers are often engaged in efforts to incentivize face-to-face visits. This is especially true the case in the post-pandemic era when providers are refocusing their attention on in-person visits. We discussed several possible forms of such incentives in Section 1: improving the in-clinic visit experience, investing in a better physical service environment, issuing a transportation bonus, and so on. In this section, we explore the impact of these efforts on the provider and patients. We will use the term bonus as an umbrella term for such incentives. Let b denote the bonus for an in-person visit. The patient utility functions then become

$$\begin{aligned} U_f^b(\lambda_f,\lambda_v) &= R - T + b - \theta_f\left[\mu_f-\lambda_f-\delta\lambda_v\right]^{-1}; \\ U_v^b(\lambda_f,\lambda_v) &= R - \theta_v\left[\mu_v-\lambda_v\right]^{-1} - \delta T + \delta b - \delta\theta_f\left[\mu_f-\lambda_f-\delta\lambda_v\right]^{-1}. \end{aligned} \tag{20}$$

To limit the patient's bonus so as not to exceed the payer's reward, we require that $b \le r_f$. It follows that $\delta b \le r_v$. These stipulations ensure that the expected net revenue is positive.

The effect of the bonus on patients' choice is not straightforward. On the one hand, the bonus might encourage patients to use the face-to-face channel for their first visit. On the other hand, the bonus might encourage them to use the virtual channel, and in turn, benefit from the bonus if and when a supplementary in-clinic visit is required. We start by studying patients' equilibrium given a fixed bonus. Then, we characterize the optimal joint bonus and capacity decision that maximizes the provider's revenue. Finally, we examine the impact of the bonus on the total access rate, i.e., the proportion of patients who do not balk. The total access rate can be viewed a measure of social welfare (Cakici and Mills 2022).

5.1. Patients' Equilibrium Strategy

Similarly to the basic model that does not offer a bonus (Section 4.1), given a fixed bonus b, there could be seven types of patient strategies: B, V, F, BVF, BF, BV and VF. Following Theorem 1, Corollary 1 summarizes patients' equilibrium strategies for any service capacity allocation (μ_f, μ_v) and non-negative bonus $b \geq 0$.

COROLLARY 1.

- *When $(\mu_f, \mu_v) \in$ Region B, there exists a unique equilibrium strategy – strategy B – such that $\lambda_f = \lambda_v = 0$. Region B satisfies the following conditions:*

$$\theta_v \mu_v^{-1} + \delta\theta_f \mu_f^{-1} \geq R - \delta T + \delta b;$$
$$\mu_f \leq \theta_f \left[R - T + b\right]^{-1}.$$

- *When $(\mu_f, \mu_v) \in$ Region V, there exists a unique equilibrium strategy – strategy V – such that $\lambda_v = \lambda$. Region V satisfies the following conditions:*

$$\theta_v \left[\mu_v - \lambda\right]^{-1} - (1-\delta)\theta_f \left[\mu_f - \delta\lambda\right]^{-1} \leq (1-\delta)(T - b);$$
$$\theta_v \left[\mu_v - \lambda\right]^{-1} + \delta\theta_f \left[\mu_f - \delta\lambda\right]^{-1} \leq R - \delta T + \delta b;$$
$$\mu_f \geq \delta\lambda + \delta\theta_f \left[R - \delta T + \delta b\right]^{-1};$$
$$\mu_v \geq \lambda + \theta_v \left[R - \delta T + \delta b\right]^{-1}.$$

- *When $(\mu_f, \mu_v) \in$ Region F, there exists a unique equilibrium strategy – strategy F – such that $\lambda_f = \lambda$. Region F satisfies the following conditions:*

$$\mu_f \geq \lambda + \theta_f \left[R - T + b\right]^{-1};$$
$$\theta_v \mu_v^{-1} - (1-\delta)\theta_f \left[\mu_f - \lambda\right]^{-1} \geq (1-\delta)(T - b).$$

- *When $(\mu_f, \mu_v) \in$ Region BVF, there exists a unique equilibrium strategy – strategy BVF – such that $\lambda_f + \lambda_v \leq \lambda$. Region BVF satisfies the following conditions:*

$$\mu_v \geq \theta_v \left[(1-\delta)R\right]^{-1},$$
$$\mu_f - \delta\mu_v \geq \theta_f \left[R - T + b\right]^{-1} - \delta\theta_v \left[(1-\delta)R\right]^{-1};$$
$$\mu_f + (1-\delta)\mu_v \leq \lambda + \theta_f \left[R - T + b\right]^{-1} + \theta_v R^{-1}.$$

The effective arrival rates (λ_f, λ_v) can be derived from $U_f^b(\lambda_f, \lambda_v) = U_v^b(\lambda_f, \lambda_v) = 0$.

- *When $(\mu_f, \mu_v) \in$ Region BF, there exists a unique equilibrium strategy – strategy BF – such that $\lambda_f \leq \lambda,\ \lambda_v = 0$. Region BF satisfies the following conditions:*

$$\begin{aligned}&\theta_f\left[R-T+b\right]^{-1} \leq \mu_f \leq \lambda + \theta_f\left[R-T+b\right]^{-1};\\&\mu_v \leq \theta_v\left[(1-\delta)R\right]^{-1}.\end{aligned}$$

The effective arrival rate λ_f can be be derived from $U_f^b(\lambda_f, 0) = 0$.

- *When $(\mu_f, \mu_v) \in$ Region BV, there exists a unique equilibrium strategy – strategy BV – such that $\lambda_f = 0,\ \lambda_v \leq \lambda$. Region BV satisfies the following conditions:*

$$\begin{aligned}&\theta_v\mu_v^{-1} + \delta\theta_f\mu_f^{-1} \leq R - \delta T + \delta b,\\&\theta_v\left[\mu_v - \lambda\right]^{-1} + \delta\theta_f\left[\mu_f - \delta\lambda\right]^{-1} \geq R - \delta T + \delta b \quad \text{for } \mu_v > \lambda \text{ and } \mu_f > \delta\lambda;\\&\mu_f - \delta\mu_v \leq \theta_f\left[R-T+b\right]^{-1} - \delta\theta_v\left[(1-\delta)R\right]^{-1}.\end{aligned}$$

The effective arrival rate λ_v can be derived by $U_v^b(0, \lambda_v) = 0$.

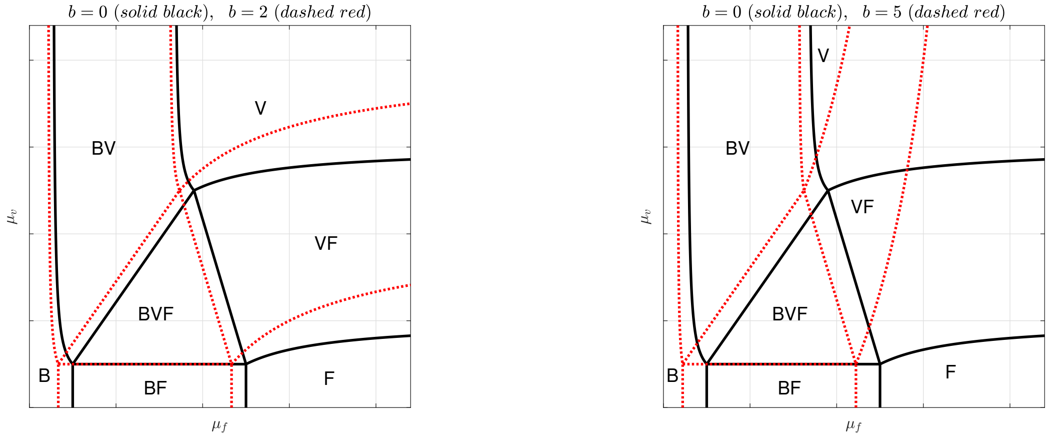
- *When $(\mu_f, \mu_v) \in$ Region VF, there exists a unique equilibrium strategy – strategy VF – such that $\lambda_f + \lambda_v = \lambda$. Region VF satisfies the following conditions:*

$$\begin{aligned}&\theta_v\mu_v^{-1} - (1-\delta)\theta_f\left[\mu_f - \lambda\right]^{-1} \leq (1-\delta)(T-b) \quad \text{for } \mu_f > \lambda,\\&\theta_v\left[\mu_v - \lambda\right]^{-1} - (1-\delta)\theta_f\left[\mu_f - \delta\lambda\right]^{-1} \geq (1-\delta)T \quad \text{for } \mu_v > \lambda,\\&\mu_f + (1-\delta)\mu_v \geq \lambda + \theta_f\left[R-T+b\right]^{-1} + \theta_v R^{-1};\\&\mu_f \geq \delta\lambda + \theta_f\left[R-T+b\right]^{-1}.\end{aligned}$$

The effective arrival rates (λ_f, λ_v) can be derived by $\lambda_f + \lambda_v = \lambda$ and $U_f^b(\lambda_f, \lambda_v) = U_v^b(\lambda_f, \lambda_v)$.

Per Corollary 1, patients' equilibrium strategies exhibit similar structures as the one without a bonus. The divisions of these regions, however, slightly changes as the utility functions incorporate b. Note that the sign of $(1-\delta)(T-b)$ affects the boundary shapes between Region V and Region VF and between Region F and Region VF. Figure 4 illustrates these changes in the regions' structures for different values of b. We can see that with a positive bonus b, the boundaries are moving to the left, resulting in a smaller Region B. When $T-b$ is positive (the left plot), the shapes of the boundaries remain the same. When, however, $T-b$ is negative (the right plot), the shapes change: Region V, where all patients choose the virtual channel, shrinks significantly, while Region F, where all patients choose the face-to-face channel, grows. Generally speaking, providing a bonus for in-person visits can increase service utilization, especially for the face-to-face channel.

Figure 4 **Equilibrium regions for different values of b.**



$b=0$ (*solid black*), $b=2$ (*dashed red*)

$b=0$ (*solid black*), $b=5$ (*dashed red*)

5.2. Optimal Joint Incentive and Capacity Decision

In this section, we study the provider's joint decision regarding the bonus b and the capacity allocation (μ_f, μ_v), to maximize the expected net reward

$$(r_f - b)\lambda_f + (r_v - \delta b)\lambda_v,$$

where (λ_f, λ_v) is the anticipated effective arrival rates in equilibrium given a bonus b and capacity allocation (μ_f, μ_v). The challenge here is the fact that different b's will lead to different feasible equilibrium regions, and thus different forms of the optimal capacity allocation, making joint optimization challenging.

In the following analysis, we first study the optimal capacity allocation given any b. Conveniently, because we replace T with $T-b$, all the results follow directly from our analysis in Section 4.2. To simplify the analysis, in Corollary 2, we summarize all possible scenarios for the optimal capacity allocation for any given b.

Corollary 2. *For any given b, the optimal capacity allocation $(\mu_f^*(b), \mu_v^*(b))$ must result in one of the following scenarios:*

1. ***PURE Scenarios**: All patients choose a pure strategy:*
   - ***The PURE-B Scenario**: All patients choose to balk;*
   - ***The PURE-F Scenario**: All patients choose the face-to-face channel;*
   - ***The PURE-V Scenario**: All patients choose the virtual channel.*
2. ***The NV Scenario**: No patient chooses the virtual channel. Therefore, $\lambda_v = 0$ and*

$$\lambda_f = \min\left\{\lambda, \mu - \theta_f \left[R - T + b\right]^{-1}\right\}.$$

3. ***The*** *NF* ***Scenario****: No patient chooses the face-to-face channel. Therefore, $\lambda_f = 0$ and $\lambda_v = \lambda_v^{NF}$, where*

$$\lambda_v^{NF} = \begin{cases} \min\left\{\lambda, \frac{\mu}{1+\delta} - \frac{(\sqrt{\delta\theta_f}+\sqrt{\theta_v})^2}{(1+\delta)(R-\delta T+\delta b)}\right\}, & \textit{if } \frac{\delta(\theta_f-\theta_v)+(1-\delta)\sqrt{\delta\theta_f\theta_v}}{(R-\delta T+\delta b)} \leq \frac{\theta_f}{R-T+b} - \frac{\delta\theta_v}{(1-\delta)R}, \\ \frac{\mu}{1+\delta} - \frac{\theta_f}{(1+\delta)(R-T+b)} - \frac{\theta_v}{(1+\delta)(1-\delta)R}, & \textit{otherwise.} \end{cases} \tag{21}$$

4. ***The*** *NB* ***Scenario****: No patient balks. Therefore, $\lambda_v = \lambda_v^{NB}$, where*

$$\lambda_v^{NB} = \tfrac{1}{\delta}\left(\mu - \lambda - \theta_v\left[(1-\delta)R\right]^{-1} - \theta_f\left[R-T+b\right]^{-1}\right) \quad \textit{and} \quad \lambda_f = \lambda - \lambda_v^{NB}. \tag{22}$$

These six scenarios constitute a refined version of the regions defined in Theorem 1, as they represent the possible cases of optimal capacity allocation. Specifically, Scenario NV represents the optimal allocation in Region BF as well as the special case in Region F where $\mu > \theta_f/(R-T+b)$. Scenario NF represents the optimal allocation in Region BV as well as the special case in Region V. Scenario NB represents the optimal allocation in Region VF, excluding the cases of Region V or F.

Next, we determine the optimal b by studying how $(\mu_f^*(b), \mu_v^*(b))$ change with b. In particular, we identify the optimal b within each scenario. Then, we obtain the global optimal joint decision of b^* and (μ_f^*, μ_v^*) by comparing these six scenarios.

For a $PURE$ scenario, the optimal b is straightforward – in all the scenario variations, there is no need to consider a bonus with restrictions because, as demonstrated later, the optimal b in these scenarios would be zero. For the other scenarios, we first need to determine the b region that can make a specific scenario possible; only then we can ascertain the optimal b. Recall that we require $b \leq r_f$ and hence $b < r_v/\delta$.

Let $\underline{b}_i$, $i = NV, NF, NB$, denote the smallest b that enables achieving Scenario i. Specifically,

- The smallest b that enables achieving Scenario NV is

$$\underline{b}_{NV} = \theta_f\mu^{-1} - R + T. \tag{23}$$

- The smallest b that enables achieving Scenario NF is

$$\underline{b}_{NF} = \tfrac{1}{\delta}\left(\theta_v\Phi^{-1}(\mu) + \delta\theta_f\left[\mu - \Phi(\mu)\right]^{-1} - R + \delta T\right), \tag{24}$$

where $\Phi(\mu)$ is defined in (19).

- The smallest b that enables achieving Scenario NB is

$$\underline{b}_{NB} = \begin{cases} \infty, & \text{if } \mu \leq \lambda + \theta_v \left[(1-\delta)R\right]^{-1}, \\ T - R + \theta_f \left[\mu - \lambda - \theta_v \left[(1-\delta)R\right]^{-1}\right]^{-1}, & \text{otherwise.} \end{cases} \tag{25}$$

Note that $\underline{b}_{NV}$, $\underline{b}_{NF}$ and $\underline{b}_{NB}$ function as the lower bound on b when studying the optimal bonus for a specific scenario. Lemma 1 determines a global upper bound for the optimal bonus b.

Lemma 1. *The optimal bonus b^* must satisfy $b^* \leq \bar{b}$ where*

$$\bar{b} = \begin{cases} \infty, & \textit{if } \mu \leq (1+\delta)\lambda + \theta_v \left[(1-\delta)R\right]^{-1}, \\ T - R + \theta_f \left[\mu - (1+\delta)\lambda - \theta_v \left[(1-\delta)R\right]^{-1}\right]^{-1}, & \textit{otherwise.} \end{cases} \tag{26}$$

In other words, $\bar{b}$ is the smallest b that enables the system to perform as a "large" system, as defined in Definition 1. In particular, when b hits $\bar{b}$, the system is able to attract all patients under the optimal capacity allocation. Therefore, there is no need to consider $b > \bar{b}$.

After characterizing the feasible region of b, we are now ready to analyze the optimal b in each scenario described in Corollary 2. Proposition 4 summarizes the results.

Proposition 4. *The optimal bonus for an in-person visit for each scenario is*

- ***Any PURE Scenario:*** $b^{PURE*} = 0$.
- ***The NV Scenario:***

$$b^{NV*} = \underset{b \in [\underline{b}^+_{NV},\ r_f]}{\arg\max}\ (r_f - b) \min\left\{\mu - \theta_f \left[R - T + b\right]^{-1},\ \lambda\right\}. \tag{27}$$

In particular, when $[\underline{b}^+_{NV},\ r_f] \neq \emptyset$,

$$b^{NV*} = \min\left\{\max\left\{\underline{b}^+_{NV},\ \sqrt{\theta_f(R - T + r_f)\mu^{-1}} - R + T\right\},\ r_f,\ \bar{b}_{NV}\right\},$$

where $\bar{b}_{NV} = \infty$ if $\lambda \geq \mu$; otherwise, $\bar{b}_{NV} = (\theta_f \left[\mu - \lambda\right]^{-1} - R + T)^+$.

- ***The NF Scenario:***

$$b^{NF*} = \underset{b \in [\underline{b}^+_{NF},\ \min\{r_v/\delta,\ \bar{b}\}]}{\arg\max}\ (r_v - \delta b)\lambda_v^{NF}, \tag{28}$$

where λ_v^{NF} is defined in (21). A closed-form expression for b^{NF*} is developed in Appendix A under an additional technical assumption.

- ***The*** NB ***Scenario:***

$$b^{NB*} = \underset{b\in[\underline{b}^{+}_{NB},\ \min\{r_f,\ r_v/\delta,\ \bar{b}\}]}{\arg\max} (r_f - b)\lambda + [r_v - r_f + (1-\delta)b]\,\lambda_v^{NB}, \tag{29}$$

where λ_v^{NB} *is defined in* (22). *In particular, when* $[\underline{b}^{+}_{NB},\ \min\{r_f,\ r_v/\delta,\ \bar{b}\}] \neq \emptyset$*:*

$$b^{NB*} = \begin{cases} \underline{b}^{+}_{NB} \text{ or } \min\{r_f,\ r_v/\delta,\ \bar{b}\}, & \text{if } r_v - r_f - (1-\delta)(R-T) \leq 0 \\ \min\{\max\{\underline{b}^{+}_{NB}, b^{NB\#}\}, r_f,\ r_v/\delta,\ \bar{b}\}, & \text{otherwise,} \end{cases}$$

where

$$b^{NB\#} = \begin{cases} \infty, & \text{if } (1-\delta)(\mu-\lambda) - \delta\lambda - \theta_v/R \geq 0 \\ \sqrt{\frac{\theta_f(r_v - r_f - (1-\delta)(R-T))}{-(1-\delta)(\mu-\lambda)+\delta\lambda+\theta_v/R}} - R + T, & \text{otherwise.} \end{cases}$$

After characterizing the optimal bonus for each scenario, we identify the one with the largest expected net reward as the global optimal bonus.

The following analysis narrows the options for the optimal scenario under a given total service capacity μ. Specifically, we prescribe what scenario as defined in Corollary 2 and Proposition 4 can be optimal depending the size of the system, i.e., μ. We start by defining a few useful thresholds for μ. Let $\underline{\mu}_i$, $i = NV, NF, NB$, denote the smallest capacity required to assure Scenario i. Specifically,

- The smallest capacity required to assure Scenario NV is defined as

$$\underline{\mu}_{NV} = \theta_f\,[R - T + r_f]^{-1}.$$

- The smallest capacity required to assure Scenario NF, $\underline{\mu}_{NF}$ is achieved by solving the following equation for μ:

$$R - \delta(T - r_v/\delta) - \theta_v\Phi^{-1}(\mu) - \delta\theta_f\,[\mu - \Phi(\mu)]^{-1} = 0,$$

where $\Phi(\mu)$ is defined in (19).

- The smallest capacity required to assure Scenario NB:

$$\underline{\mu}_{NB} = \lambda + \theta_f\left[R - T + \min\{r_f, r_v\delta^{-1}\}\right]^{-1} + \theta_v\,[(1-\delta)R]^{-1}.$$

We also define

$$\bar{\mu} = (1+\delta)\lambda + \theta_f\,[R-T]^{-1} + \theta_v\,[(1-\delta)R]^{-1}.$$

When $\mu \geq \bar{\mu}$, the provider is able to achieve the $PURE-V$ and $PURE-F$ scenarios with $b = 0$. In other words, $\bar{\mu}$ is a bonus-consideration threshold.

We are now ready for Proposition 5, which establishes the optimal scenario for each μ.

PROPOSITION 5. *The optimal scenario for a given service capacity μ is:*

1. *When* $\mu \le \min\{\underline{\mu}_{NV}, \underline{\mu}_{NF}\}$, *Scenario* $PURE - B$ *is optimal.*
2. *When* $\min\{\underline{\mu}_{NV}, \underline{\mu}_{NF}\} < \mu < \max\{\underline{\mu}_{NV}, \underline{\mu}_{NF}\}$:
   - *If* $\underline{\mu}_{NV} < \underline{\mu}_{NF}$, *Scenario* NF *is optimal.*
   - *If* $\underline{\mu}_{NV} > \underline{\mu}_{NF}$, *Scenario* NV *is optimal.*
3. *When* $\max\{\underline{\mu}_{NV}, \underline{\mu}_{NF}\} \le \mu < \underline{\mu}_{NB}$, *one of the following two scenarios is optimal:* NV *or* NF.
4. *When* $\underline{\mu}_{NB} \le \mu < \bar{\mu}$, *one of the following three scenarios is optimal:* NV, NF *or* NB.
5. *When* $\mu \ge \bar{\mu}$, *one of the following two scenarios is optimal:* $PURE - F$ *or* $PURE - V$.

Note that the optimal bonus is zero in all the $PURE$ scenarios. That is, the bonus is effective only when the system is neither too small nor too large, i.e., when $\min\{\underline{\mu}_{NV}, \underline{\mu}_{NF}\} < \mu < \bar{\mu}$. Particularly, if $r_f \ge r_v$, we can arrive at Corollary 3:

COROLLARY 3. *If* $r_f \ge r_v$, $b^* = 0$ *when* $\mu \ge \lambda + \theta_f [R - T]^{-1}$.

5.3. The Incentive's Impact on the Access Rate

The adoption of telemedicine maintained and improved access to care during the pandemic. With the pandemic potentially drawing to a close, providers are starting to reengage patients in face-to-face visits. Providers can offer a bonus for in-person visits to increase their revenue. The impact of the bonus on patient access to care, however, is not clear. In our model, patient access to care is operationalized by the total effective arrival rate of patients who access the healthcare service (and do not balk); specifically, $\lambda_f^* + \lambda_v^*$, where λ_f^* and λ_v^* are the effective arrival rates in equilibrium under the provider's optimal joint decision b^* and (μ_f^*, μ_v^*). If such a bonus, which increases the provider's revenue, decreases the total access rate, then its adoption deserves caution.

Proposition 6 identifies the sufficient conditions that assure that the total access rate will not decrease when utilizing a bonus.

PROPOSITION 6. *Let* $\lambda_f^*(b)$ *and* $\lambda_v^*(b)$ *denote the effective arrival rates in equilibrium under bonus* b *and the corresponding optimal capacity allocation* $(\mu_f^*(b), \mu_v^*(b))$. *Let* b^* *denote the optimal bonus that achieves maximum revenue. Fixing all model parameters, the total access rate will not decrease when utilizing a bonus (i,e,* $\lambda_f^*(b^*) + \lambda_v^*(b^*) \ge \lambda_f^*(0) + \lambda_v^*(0)$*), if* $\mu > \theta_f [R - T]^{-1} + \theta_v [(1-\delta)R]^{-1}$ *and*

$$\frac{r_v}{r_f} \le \min\left\{1, (1+\delta)\left(\mu - \theta_f [R-T]^{-1}\right)\left[\mu - \left(\sqrt{\delta\theta_f} + \sqrt{\theta_v}\right)^2 [R - \delta T]^{-1}\right]^{-1}\right\} \tag{30}$$

or

$$\frac{r_v}{r_f} \geq (1+\delta)\left(\mu - \theta_f\left[R-T\right]^{-1}\right)\left[\mu - \theta_f\left[R-T\right]^{-1} - \theta_v\left[(1-\delta)R\right]^{-1}\right]^{-1} > 1. \tag{31}$$

Per Proposition 6, we are interested in the situation where the total service capacity is not too small. When the ratio between r_v and r_f is small or large enough, the total access rate will not decrease when adopting the optimal bonus. To see this, let us consider the case where r_v/r_f is small enough, i.e., (30) holds. In this case, the face-to-face channel is preferable. Thus, the provider will tend not to utilize the virtual channel and we will have $\lambda_v^*(0) = 0$. After utilizing the optimal bonus, the provider achieves higher revenue, which implies a larger access rate because $r_f \geq r_v$. More specifically, we note that $r_f\lambda_f^*(b^*) + r_f\lambda_v^*(b^*) \geq r_f\lambda_f^*(b^*) + r_v\lambda_v^*(b^*) \geq r_f\lambda_f^*(0)$. A similar argument can explain why when r_v/r_f is large enough, the total access rate will not decrease when the optimal bonus is utilized.

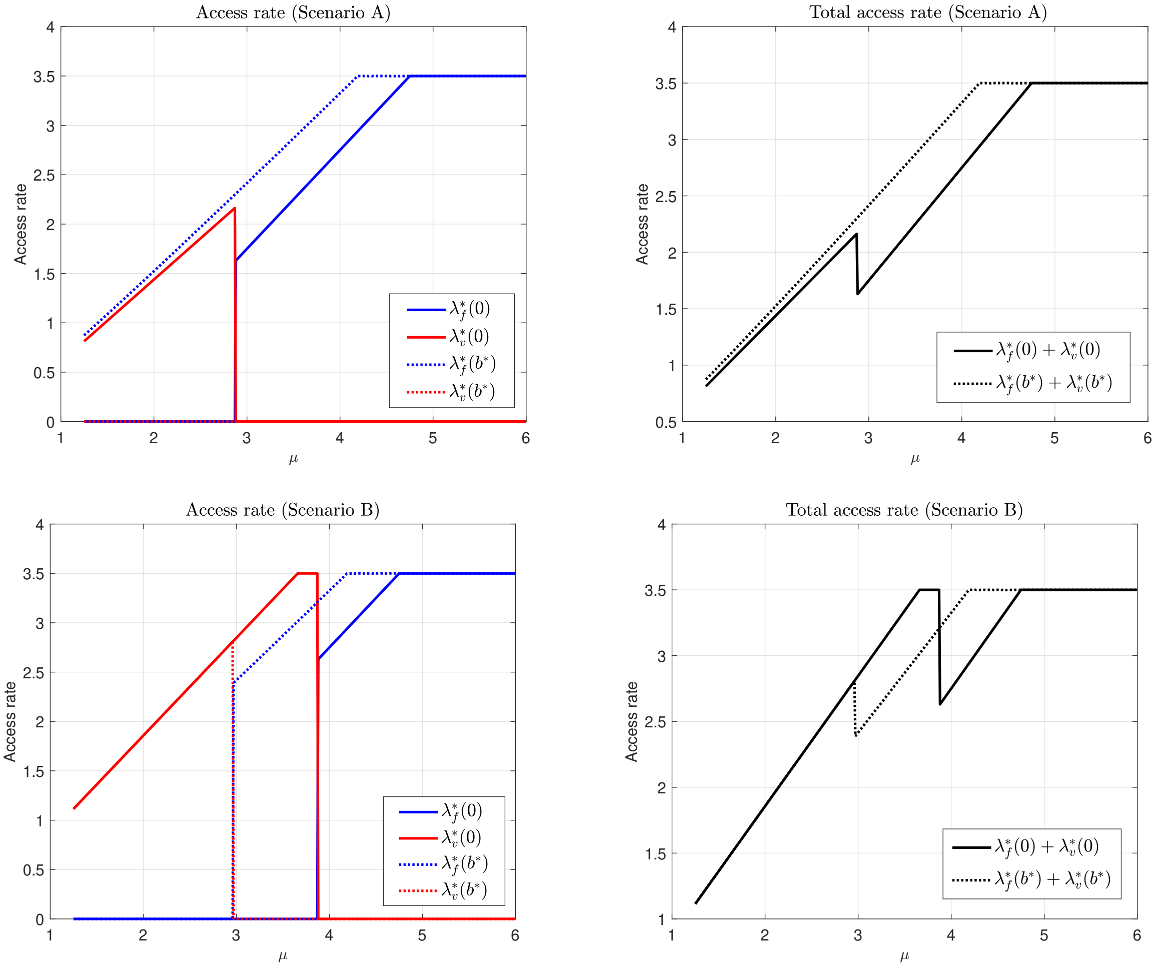
Figure 5 plots how arrival rates in equilibrium under the provider's optimal decision change with the total service capacity μ with and without adopting the bonus, for a given set of model parameters. Specifically, Scenario A in Figure 5 ($r_v/r_f = 0.75$ and $\delta = 0.2$) demonstrates the case where the bonus does not hurt the total access rate. We observe that when utilizing a bonus, the arrival rate curve shifts to the left. That is, what the bonus effectively does is to increase the system's capacity, which allows it to attract more patients and generate higher revenue. Note that when μ is large, the bonus effect diminishes since systems with large capacities do not need a bonus to attract all patients. The provider's increased revenue, however, does not always go hand in hand with increased social welfare. We summarize the cases where the total access rate may decrease when utilizing the optimal bonus:

1. When $r_v \geq r_f$: $\lambda_f^*(0) > 0$, $\lambda_f^*(b^*) = 0$ and $\lambda_v^*(b^*) < \lambda$.
2. When $r_f \geq r_v$: $\lambda_v^*(0) > 0$, $\lambda_v^*(b^*) = 0$ and $\lambda_f^*(b^*) < \lambda$.

In these two cases, patient utility increases with the bonus; consequently, the provider can increase their revenue by attracting fewer patients to the more "profitable" service channel, which generates a higher revenue.

Scenario B in Figure 5 ($r_v/r_f = 0.75$ and $\delta = 0.01$) illustrates an example where the access rate actually decreases when the provider adopts the optimal bonus. When $\mu \approx 3$–4, the equilibrium without a bonus is in Region BV, whereas with a bonus, it shifts to Region BF, meaning that the provider opts to close the virtual channel and open the face-to-face channel only because the latter becomes more appealing to patients with a

Figure 5 **Access rate comparison as a function of μ – with and without a bonus for in-person visits.**



bonus. The provider can actually achieve higher revenue with fewer patient visits through the face-to-face channel – which, in turn, hurts the total access rate.

To conclude, if the payment ratio between two channels is significant, the social welfare, captured by total patient access rate, will not be hurt when the revenue-chasing provider offers a bonus. Otherwise, the bonus needs to be used with caution, especially when the optimal system design that utilizes the bonus shifts to the channel with a higher payment rate and not all patients can be served, i.e., there is balking.

6. Discussion and Concluding Remarks

Motivated by the rise of telemedicine in healthcare practice, we study an outpatient care provider who can serve patients through a face-to-face channel and/or a virtual channel. The virtual channel can reduce the patient's burden and associated cost involved in physical travel. It cannot, however, resolve all patient care needs and, thus, a supplementary in-person visit may be required. A revenue-driven provider has two operational levers to influence patients' strategic choices – allocating capacity between the two service channels and providing incentives for in-person visits. We develop a stylized queueing model to study patients' choices and the provider's optimal decisions.

The capacity coordination between the two channels needs to be carefully balanced. Increasing offline capacity can attract more patients to the face-to-face channel, whereas increasing online capacity may not attract more patients to the virtual channel because more offline capacity may be needed to support supplementary in-person visits. Despite the growing popularity of telemedicine, offering online service may not be the best choice for all providers. It turns out that the total service capacity available to the provider plays a pivotal role in how the capacity should be allocated between the two channels. The provider with small or large capacity is better off focusing on one channel to achieve the highest revenue; sometimes focusing on the in-person channel can be the right choice. In contrast, the medium-sized provider may need to run both channels simultaneously.

The incentive for in-person visits improves patient utilities vis-à-vis channels and can serve as a useful lever to attract in-patient visits (and boost revenues). For the provider, whether to initiate the incentive also depends on the available service capacity. Providers with a small- or large-sized service capacity have no need to use incentives, but a provider with medium-sized capacity may benefit from it.

Our analysis highlights the impact in-person incentives can have on patient access to care. In some cases, using such incentives, the provider can achieve higher revenue by attracting fewer patients to the service channel where compensation is higher – leading to a decrease in total access rate. Nevertheless, when the payment gap between the two channels is large enough, the increased patient utilities due to in-person incentives will help the provider attract more visits, and thus both the provider's revenue and the total access rate will increase. To put this discussion in the context of a concrete reimbursement regime, consider a fee-for-service model. Recall that $r_f = r_f^{\text{fee}}$ and $r_v = r_v^{fee} + \delta r_f^{\text{fee}}$ in Remark 1. Condition (30) implies that $r_v^{\text{fee}} \leq (1-\delta) r_f^{\text{fee}}$, meaning that if the reimbursement for a single and sole virtual visit is sufficiently small than that for in-person visit, the adoption of in-person incentives would not hurt overall patient access to care. This implication supports the ongoing policy discussion that advocates payment equity rather than parity for telemedicine (Shachar et al. 2020). In the post-pandemic era with more providers likely to invest in efforts to attract in-patient visits, proper financial incentives by the payer can be useful to ensure patient access to care is not negatively affected.

There are several interesting directions for future research. First, from the operational perspective, a different operational lever – patient prioritization – may be considered to increase revenue and social welfare. The prioritization might drive patients to choose one channel over the other and it would be interesting to study the conditions under which it is beneficial to do so. Another possible direction is to consider a setting where returning patients are served in a separate channel with dedicated capacity and to study the optimal design of such systems. Lastly, although we interpret some of our research findings through the lens of reimbursement policy, the design of reimbursement policies is not the focus of this work and would be a fruitful direction to explore in the future.

References

Ashwood, J.S., A. Mehrotra, D. Cowling, L. Uscher-Pines. 2017. Direct-to-consumer telehealth may increase access to care but does not decrease spending. *Health Affairs* **36**(3) 485–491.

Baron, O., O. Berman, D. Krass, J. Wang. 2014. Using strategic idleness to improve customer service experience in service networks. *Operations Research* **62**(1) 123–140.

Baron, O., X. Chen, Y. Li. 2022. Omnichannel services: The false premise and operational remedies. *Management Science* .

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

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

Bokolo, A.J. 2020. Use of telemedicine and virtual care for remote treatment in response to COVID-19 pandemic. *Journal of Medical Systems* **44**(7) 1–9.

Cakici, O.E., A.F. Mills. 2022. Telehealth in acute care: Pay parity and patient access. *Working paper* .

Cordine, J., J. Fowkes, R. Malani, L. Medford-Davis. 2022. Patients love telehealth—physicians are not so sure. *McKinsey & Company. Patients love telehealth—physicians are not so sure* .

Green, L.V., S. Savin. 2008. Reducing delays for medical appointments: A queueing approach. *Operations Research* **56**(6) 1526–1538.

Hassin, R., M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*, vol. 59. Springer Science & Business Media.

Hassin, R., R. Roet-Green. 2020. On queue-length information when customers travel to a queue. *Manufacturing & Service Operations Management* **23**(4) 989–1004.

Huang, Fengfeng, Pengfei Guo, Yulan Wang. 2022. Modeling patients' illness perception and equilibrium analysis of their doctor shopping behavior. *Production and Operations Management* **31**(3) 1216–1234.

Hur, J., M.C. Chang. 2020. Usefulness of an online preliminary questionnaire under the COVID-19 pandemic. *Journal of Medical Systems* **44** 1–2.

Kadir, M.A. 2020. Role of telemedicine in healthcare during COVID-19 pandemic in developing countries. *Telehealth and Medicine Today* .

Li, K.Y., Z. Zhu, S. Ng, C. Ellimoottil. 2021. Direct-to-consumer telemedicine visits for acute respiratory infections linked to more downstream visits: Study examines the association between telemedicine and downstream health care utilization. *Health Affairs* **40**(4) 596–602.

Liu, Nan, Willem van Jaarsveld, Shan Wang, Guanlian Xiao. 2023. Managing outpatient service with strategic walk-ins. *Management Science* .

McConnochie, K.M., S.D. Ronis, N.E. Wood, P.K. Ng. 2015. Effectiveness and safety of acute care telemedicine for children with regular and special healthcare needs. *Telemedicine and e-Health* **21**(8) 611–621.

Monaghesh, E., A. Hajizadeh. 2020. The role of telehealth during COVID-19 outbreak: A systematic review based on current evidence. *BMC Public Health* **20**(1) 1–9.

Naor, P. 1969. The regulation of queue size by levying tolls. *Econometrica* 15–24.

O'Reilly-Jacob, M., P. Mohr, M. Ellen, C. Petersen, C. Sarkisian, S. Attipoe, E. Rich. 2021. Digital health & low-value care. *Healthcare*, vol. 9. Elsevier, 100533.

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

Reed, M., J. Huang, I. Graetz, E. Muelly, A. Millman, C. Lee. 2021. Treatment and follow-up care associated with patient-scheduled primary care telemedicine and in-person visits in a large integrated health system. *JAMA network open* **4**(11) e2132793–e2132793.

Rosenthal, E. 2021. Telemedicine is a tool, not a replacement for your doctor's touch. *New York Times* .

Shachar, Carmel, Jaclyn Engel, Glyn Elwyn. 2020. Implications for telehealth in a postpandemic future: regulatory and privacy issues. *Jama* **323**(23) 2375–2376.

Shi, Z., A. Mehrotra, C.A. Gidengil, S.J. Poon, L. Uscher-Pines, K.N Ray. 2018. Quality of care for acute respiratory infections during direct-to-consumer telemedicine visits for adults. *Health Affairs* **37**(12) 2014–2023.

The American Academy of Family Physicians. 2019. How to legally provide free transportation to patients. Accessed Feb 14, 2023: https://www.aafp.org/pubs/fpm/blogs/inpractice/entry/patient_inducements.html.

Wong, M.Y.Z., D.V. Gunasekeran, S. Nusinovici, C. Sabanayagam, K. K. Yeo, C.-Y. Cheng, Y.-C. Tham. 2021. Telehealth demand trends during the COVID-19 pandemic in the top 50 most affected countries: Infodemiological evaluation. *JMIR public health and surveillance* **7**(2) e24445.

Zacharias, C., M. Armony. 2016. Joint panel sizing and appointment scheduling in outpatient care. *Management Science* **63**(11) 3978–3997.

Zychlinski, N. 2023. Managing queues with reentrant customers in support of hybrid healthcare. *Stochastic Systems, forthcoming* .

Online Appendix

Appendix A: A Closed-Form Expression for b^{NF*}

The optimal bonus for Scenario NF can be expressed in a closed form by making the following technical assumption on the utility functions. The assumption restricts the return probability δ from being too large. That is aligned with the practical return probability, which is estimated to be 6–20% (Yamamoto 2014, Uscher-Pines et al. 2016, Shi et al. 2018).

Assumption A.1. *The following relation holds*

$$\frac{\delta(\theta_f-\theta_v)+(1-\delta)\sqrt{\delta\theta_f\theta_v}}{R-\delta T}<\frac{\theta_f}{R-T}-\frac{\delta\theta_v}{(1-\delta)R}.$$

Proposition A.1. *Under Assumption A.1, when* $[(\underline{b}_{NF})^+,\ \min\{r_v./\delta,\ \bar{b}\}]\neq\emptyset$,

$$b^{NF*}=\min\left\{\max\left\{b^{NF\#},\ (\underline{b})^+\right\},\ r_v/\delta,\ \bar{b}\right\},$$

where

$$b^{NF\#}=\begin{cases}b^1_{NF}, & \textit{if } b^1_{NF}\le b^0\\ b^2_{NF}, & \textit{otherwise.}\end{cases}$$

Here b^0 *solves*

$$\frac{\delta(\theta_f-\theta_v)+(1-\delta)\sqrt{\delta\theta_f\theta_v}}{R-\delta T+\delta b^0}=\frac{\theta_f}{R-T+b^0}-\frac{\delta\theta_v}{(1-\delta)R}.$$

b^1_{NF} *and* b^2_{NF} *are the optimal solutions to the two cases of* $(r_v-\delta b)\lambda_v^{NF}(b)$. *Specifically,*

$$b^1_{NF}=\min\left\{\frac{(\sqrt{\delta\theta_f}+\sqrt{\theta_v})\sqrt{R-\delta T+r_v}}{\delta\sqrt{\mu}}-\frac{R-\delta T}{\delta},\ \bar{b}^1_{NF}\right\},$$

where

$$\bar{b}^1_{NF}=\begin{cases}\infty, & \textit{if } \lambda\ge\frac{\mu}{1+\delta}\\ \frac{\left(\sqrt{\delta\theta_f}+\sqrt{\theta_v}\right)^2}{\delta(\mu-(1+\delta)\lambda)}-\frac{R-\delta T}{\delta}, & \textit{otherwise;}\end{cases}$$

$$b^2_{NF}=\min\left\{\sqrt{\frac{\theta_f(\delta R-\delta T+r_v)}{\delta\mu-\frac{\delta\theta_v}{(1-\delta)R}}}-R+T,\ \bar{b}^2_{NF}\right\},$$

where

$$\bar{b}^2_{NF}=\begin{cases}\infty, & \textit{if } \lambda\ge\frac{\mu}{1+\delta}-\frac{\theta_v}{(1+\delta)(1-\delta)R}\\ \frac{\theta_f}{\mu-(1+\delta)\lambda-\frac{\theta_v}{(1-\delta)R}}-R+T, & \textit{otherwise.}\end{cases}$$

Appendix B: Proofs of Analytical Results

Proof of Theorem 1: We conduct the analysis for each region separately:

Region *B*: Since

$$\frac{\theta_v}{\mu_v}+\frac{\delta\theta_f}{\mu_f}\geq R-\delta T \quad\text{and}\quad \mu_f\leq\frac{\theta_f}{R-T},$$

we get that $U_f(0,0)\leq 0$ and $U_v(0,0)\leq 0$ and therefore, there exists a unique equilibrium such that $\lambda_f=\lambda_v=0$.

Region *V*: Since

$$\frac{\theta_v}{\mu_v-\lambda}-\frac{(1-\delta)\theta_f}{\mu_f-\delta\lambda}\leq T-\delta T,\quad \frac{\theta_v}{\mu_v-\lambda}+\frac{\delta\theta_f}{\mu_f-\delta\lambda}\leq R-\delta T,$$
$$\mu_f\geq\delta\lambda+\frac{\delta\theta_f}{R-\delta T},\quad\text{and}\quad \mu_v\geq\lambda+\frac{\theta_v}{R-\delta T},$$

we get that $U_f(0,\lambda)\leq U_v(0,\lambda)$ and $U_v(0,\lambda)\geq 0$ and therefore, there exists a unique equilibrium such that $\lambda_f=0$ and $\lambda_v=\lambda$.

Region *F*: Since

$$\mu_f\geq\lambda+\frac{\theta_f}{R-T}\quad\text{and}\quad \frac{\theta_v}{\mu_v}-\frac{(1-\delta)\theta_f}{\mu_f-\lambda}\geq T-\delta T,$$

we get that $U_f(\lambda,0)\geq U_v(\lambda,0)$ and $U_f(\lambda,0)\geq 0$ and therefore, there exists a unique equilibrium such that $\lambda_f=\lambda$ and $\lambda_v=0$.

Region *BVF*: The equality $U_v(\lambda_f,\lambda_v)=U_f(\lambda_f,\lambda_v)=0$ yields the following conditions:

$$\lambda_v=\mu_v-\frac{\theta_v}{R-\delta R}\quad\text{and}\quad \lambda_f+\delta\lambda_v=\mu_f-\frac{\theta_f}{R-T}.$$

The constraints of Region BVF:

$$\mu_v\geq\frac{\theta_v}{(1-\delta)R},\quad \mu_f-\delta\mu_v\geq\frac{\theta_f}{R-T}-\frac{\delta\theta_v}{(1-\delta)R},\quad \mu_f+(1-\delta)\mu_v\leq\lambda+\frac{\theta_f}{R-T}+\frac{\theta_v}{R},$$

lead to $\lambda_v\geq 0$, $\lambda_f\geq 0$ and $\lambda_v+\lambda_f\leq\lambda$. Thus, there exists a unique equilibrium such that $\lambda_v+\lambda_f\leq\lambda$.

Region *BF*: The fact that $U_f(\lambda_f,0)=0$ yields the following condition:

$$\lambda_f=\min\left\{\lambda,\mu_f-\frac{\theta_f}{R-T}\right\}.$$

The constraints of Region BF:

$$\frac{\theta_f}{R-T}\leq\mu_f\leq\lambda+\frac{\theta_f}{R-T},\quad\text{and}\quad \mu_v\leq\frac{\theta_v}{(1-\delta)R}$$

lead to $0\leq\lambda_f\leq\lambda$ and $U_v(\lambda_f,0)\leq 0$. Thus, there exists a unique equilibrium such that $0\leq\lambda_f\leq\lambda$ and $\lambda_v=0$.

Region *BV*: The fact that $U_v(0,\lambda_v)=0$ yields the following conditions:

$$R-\delta T-\frac{\theta_v}{\mu_v-\lambda_v}-\frac{\delta\theta_f}{\mu_f-\delta\lambda_v}=0, \tag{B.1}$$

$$\mu_v - \lambda_v > 0 \text{ and } \mu_v - \delta\lambda_v > 0.$$

Next, we show that

1) There is only one λ_v which solves (B.1) and satisfies $\mu_f - \delta\lambda_v > 0$ and $\mu_v - \lambda_v > 0$.

2) With the constraints of Region BV, we must have $0 \le \lambda_v \le \lambda$ and $U_f(0, \lambda_v) \le 0$.

Thus, there exists a unique equilibrium such that $0 \le \lambda_v \le \lambda$ and $\lambda_f = 0$.

We start with proving 1): There is only one λ_v which solves (B.1) and satisfies $\mu_f - \delta\lambda_v > 0$ and $\mu_v - \lambda_v > 0$. Specifically,

$$\lambda_v = \frac{\mathcal{B} - \sqrt{\mathcal{D}^2 + 4\delta^2\theta_f\theta_v}}{2\mathcal{A}}, \tag{B.2}$$

where

$$\mathcal{A} = (R - \delta T)\delta, \quad \mathcal{B} = (R - \delta T)(\mu_f + \delta\mu_v) - \delta(\theta_f + \theta_v),$$

$$\mathcal{C} = (R - \delta T)\mu_f\mu_v - \theta_v\mu_f - \delta\theta_f\mu_v, \quad \mathcal{D} = (R - \delta T)(\mu_f - \delta\mu_v) - \delta(\theta_f - \theta_v).$$

Let us rewrite (B.1) as

$$(R - \delta T)\delta\lambda_v^2 - [(R - \delta T)(\mu_f + \delta\mu_v) - \delta(\theta_f + \theta_v)]\lambda_v + [(R - \delta T)\mu_f\mu_v - \theta_v\mu_f - \delta\theta_f\mu_v] = 0,$$

or alternatively, $\mathcal{A}\lambda_v^2 - \mathcal{B}\lambda + \mathcal{C} = 0$. Note that

$$\mathcal{B}^2 - 4\mathcal{A}\mathcal{C} = [(R - \delta T)(\mu_f - \delta\mu_v) - \delta(\theta_f - \theta_v)]^2 + 4\delta^2\theta_f\theta_v = \mathcal{D}^2 + 4\delta^2\theta_f\theta_v \ge 0.$$

Then, (B.1) has two possible solutions:

$$\lambda_v^{BV.1} = \frac{\mathcal{B} + \sqrt{\mathcal{D}^2 + 4\delta^2\theta_f\theta_v}}{2\mathcal{A}} \quad \text{and} \quad \lambda_v^{BV.2} = \frac{\mathcal{B} - \sqrt{\mathcal{D}^2 + 4\delta^2\theta_f\theta_v}}{2\mathcal{A}}.$$

There are two cases for $\mathcal{D}$:

- **Case 1:** $\mathcal{D} > 0$

—If $\lambda_v = \lambda_v^{BV.1}$,

$$\lambda_v^{BV.1} > \frac{\mathcal{B} + \mathcal{D}}{2\mathcal{A}} = \frac{(R - \delta T)\mu_f - \delta\theta_f}{(R - \delta T)\delta}.$$

Then, we have $\mu_f - \delta\lambda_v^{BV.1} < \frac{\delta\theta_f}{R - \delta T}$, which contradicts with either (B.1) or $\mu_f - \delta\lambda_v > 0$.

—If $\lambda_v = \lambda_v^{BV.2}$,

$$\lambda_v^{BV.2} < \frac{\mathcal{B} - \mathcal{D}}{2\mathcal{A}} = \frac{(R - \delta T)\mu_v - \theta_v}{(R - \delta T)}.$$

Then, we have $\mu_v - \lambda_v^{BV.2} > \frac{\theta_v}{R - \delta T} > 0$, and $\lambda_v^{BV.2} < \frac{\mu_f}{\delta}$ (since $\mathcal{D} > 0$), which validates (B.1), $\mu_v > \lambda_v$ and $\mu_f - \delta\lambda_v > 0$.

- **Case 2:** $\mathcal{D} \le 0$

—If $\lambda_v = \lambda_v^{BV.1}$,

$$\lambda_v^{BV.1} > \frac{\mathcal{B} - \mathcal{D}}{2\mathcal{A}} = \frac{(R - \delta T)\mu_v - \theta_v}{(R - \delta T)}.$$

Then, we have $\mu_v - \lambda_v^{BV.1} < \frac{\theta_v}{R - \delta T}$, which contradicts with (B.1) or $\lambda_v < \mu_v$.

—If $\lambda_v = \lambda_v^{BV.2}$,

$$\lambda_v^{BV.2} < \frac{\mathcal{B} + \mathcal{D}}{2\mathcal{A}} = \frac{(R - \delta T)\mu_f - \delta\theta_f}{(R - \delta T)\delta}.$$

Then, we have $\mu_f - \delta\lambda_v^{BV.2} > \frac{\delta\theta_f}{R-\delta T} > 0$, and $\lambda_v^{BV.2} < \mu_v$ (by $\mathcal{D} \leq 0$), which validates (B.1), $\mu_v > \lambda_v$ and $\mu_f - \delta\lambda_v > 0$.

In sum, $\lambda_v = \lambda_v^{BV.2}$ is the only solution which solves (B.1) and satisfies $\mu_f - \delta\lambda_v > 0$ and $\mu_v - \lambda_v > 0$.

Next, we move on to proving 2): With the constraints of Region BV, we must have $0 \leq \lambda_v \leq \lambda$ and $U_f(0, \lambda_v) \leq 0$; specifically, we prove the following:

2.1) $\boldsymbol{\lambda_v \geq 0}$: Recall the constraints of Region BV:

$$\frac{\theta_v}{\mu_v} + \frac{\delta\theta_f}{\mu_f} \leq R - \delta T, \quad \mu_f - \delta\mu_v \leq \frac{\theta_f}{R-T} - \frac{\delta\theta_v}{(1-\delta)R},$$
$$\frac{\theta_v}{\mu_v - \lambda} + \frac{\delta\theta_f}{\mu_f - \delta\lambda} \geq R - \delta T \text{ for } \mu_v > \lambda \text{ and } \mu_f > \delta\lambda.$$

Then we have $\mathcal{A} > 0$, $\mathcal{B} > 0$ and $\mathcal{C} \geq 0$. Since $\mathcal{B}^2 - 4\mathcal{A}\mathcal{C} \geq 0$, then we have $\mathcal{B} - \sqrt{\mathcal{B}^2 - 4\mathcal{A}\mathcal{C}} \geq 0$. Thus $\lambda_v \geq 0$.

2.2) $\boldsymbol{\lambda_v \leq \lambda}$: If $\mu_v \leq \lambda$ or $\mu_f \leq \delta\lambda$, then $\lambda_v < \lambda$ since $\mu_v > \lambda_v$ and $\mu_f > \delta\lambda_v$.

If $\mu_v > \lambda$ and $\mu_f > \delta\lambda$, by (B.1) and the constraint

$$\frac{\theta_v}{\mu_v - \lambda} + \frac{\delta\theta_f}{\mu_f - \delta\lambda} \geq R - \delta T \text{ for } \mu_v > \lambda \text{ and } \mu_f > \delta\lambda,$$

we must have $\lambda_v \leq \lambda$.

2.3) $\boldsymbol{U_f(0, \lambda_v) \leq 0}$: By (B.1) and

$$\mu_f - \delta\mu_v \leq \frac{\theta_f}{R-T} - \frac{\delta\theta_v}{(1-\delta)R},$$

we must have $U_f(0, \lambda_v) \leq 0$.

Region $\boldsymbol{VF}$: The fact that $\lambda_v + \lambda_f = \lambda$ and $U_f(\lambda_f, \lambda_v) = U_v(\lambda_f, \lambda_v)$ yield the following conditions

$$\frac{\theta_v}{\mu_v - \lambda_v} - \frac{(1-\delta)\theta_f}{\mu_f - \lambda + (1-\delta)\lambda_v} = (1-\delta)T, \tag{B.3}$$
$$\frac{\lambda - \mu_f}{1-\delta} < \lambda_v < \mu_v.$$

Recall the constraints of Region VF:

$$\frac{\theta_v}{\mu_v} - \frac{(1-\delta)\theta_f}{\mu_f - \lambda} \leq (1-\delta)T \text{ for } \mu_f > \lambda,$$
$$\frac{\theta_v}{\mu_v - \lambda} - \frac{(1-\delta)\theta_f}{\mu_f - \delta\lambda} \geq (1-\delta)T \text{ for } \mu_v > \lambda,$$
$$\mu_f + (1-\delta)\mu_v \geq \lambda + \frac{\theta_f}{R-T} + \frac{\theta_v}{R}, \text{ and}$$
$$\mu_f \geq \delta\lambda + \frac{\theta_f}{R-T}.$$

Since the left-hand side of (B.3) is increasing in λ_v when $(\lambda - \mu_f)/(1-\delta) < \lambda_v < \mu_v$. By the first two constraints of Region VF, there must be a unique λ_v such that $0 \leq \lambda_v \leq \lambda$ solves (B.3). By the third constraint of Region VF, it can be verified that $U_v(\lambda - \lambda_v, \lambda_v) \geq 0$. Thus, there exists a unique equilibrium such that $\lambda_f + \lambda_v = \lambda$. Specifically, solving (B.3) gives

$$\lambda_v = \frac{-\tilde{\mathcal{B}} + \sqrt{\tilde{\mathcal{D}}^2 + 4\theta_f\theta_v}}{2\tilde{\mathcal{A}}}, \tag{B.4}$$

where

$$\tilde{\mathcal{A}} = (1-\delta)T, \quad \tilde{\mathcal{B}} = (\theta_f+\theta_v)+T(\mu_f-\lambda-(1-\delta)\mu_v),$$
$$\tilde{\mathcal{C}} = \frac{\theta_v(\mu_f-\lambda)}{(1-\delta)} - T\mu_v(\mu_f-\lambda)-\theta_f\mu_v, \quad \text{and} \quad \tilde{\mathcal{D}} = (\theta_f-\theta_v)+T(\mu_f-\lambda+(1-\delta)\mu_v). \quad Q.E.D.$$

Proof of Proposition 1: Recall that in this region, $0 \le \lambda_f+\lambda_v \le \lambda$. The provider's objective function is to maximize $r_f\lambda_f + [(1-\delta)r_v+\delta r_{v,f}]\,\lambda_v$. Since in this region we have,

$$\lambda_v = \mu_v - \frac{\theta_v}{(1-\delta)R}, \quad \text{and} \quad \lambda_f = \mu_f - \delta\mu_v - \frac{\theta_f}{R-T} + \frac{\theta_v}{(1-\delta)R},$$

the objective function can be rewritten as

$$r_f\left(\mu_f - \delta\mu_v - \frac{\theta_f}{R-T} + \frac{\theta_v}{(1-\delta)R}\right) + [(1-\delta)r_v+\delta r_{v,f}]\left(\mu-\mu_f-\frac{\theta_v}{(1-\delta)R}\right),$$

which by substituting $\mu_f = \mu-\mu_v$ and removing the constants is equivalent to

$$\mu_f\left((1+\delta)r_f-(1-\delta)r_v-\delta r_{v,f}\right).$$

The healthcare provider's problem is therefore,

$$\begin{aligned}
&\max_{\mu_f} \;\; \Pi(\mu_f) = \mu_f\left((1+\delta)r_f-(1-\delta)r_v-\delta r_{v,f}\right)\\
&\text{subject to}\\
&\mu_f \le \mu - \frac{\theta_v}{(1-\delta)R};\\
&\mu_f \ge \frac{\delta}{1+\delta}\left(\mu+\frac{\theta_f}{\delta(R-T)}-\frac{\theta_v}{(1-\delta)R}\right);\\
&\mu_f \le \frac{1}{\delta}\left(\lambda+\frac{\theta_f}{R-T}+\frac{\theta_v}{R}-\mu(1-\delta)\right);\\
&\frac{\theta_f}{R-T} \le \mu_f \le \lambda+\frac{\theta_f}{R-T}.
\end{aligned} \tag{B.5}$$

The optimal solution then depends on the sign of the objective function and the boundaries of μ_f in (B.5). Q.E.D.

Proof of Proposition 2: Before characterizing the optimal capacity allocation in Region BV, we first provide an auxiliary result. Lemma B.1 proves that the unique effective arrival rate to the virtual channel (which is defined in (B.2) and here we denote it as λ_v^{BV}) is a unimodal function in μ_v, which is maximized at the unique value $\mu_v = \tilde{\mu}_v^{BV}$.

Lemma B.1. *Given (μ_f,μ_v) which satisfies (B.1) and (12)–(14), there exists a unique $\tilde{\mu}_v^{BV}$ such that $\lambda_v^{BV}(\mu-\mu_v,\mu_v)$ is increasing in μ_v when $\mu_v \le \tilde{\mu}_v^{BV}$ and $\lambda_v^{BV}(\mu-\mu_v,\mu_v)$ is decreasing in μ_v when $\mu_v \ge \tilde{\mu}_v^{BV}$.*

In Proposition 2, the results are straightforward when $\mu \ge \lambda(1+\delta)+\frac{\theta_v}{(1-\delta)R}+\frac{\theta_f}{R-T}$.

When $\mu < \lambda(1+\delta)+\frac{\theta_v}{(1-\delta)R}+\frac{\theta_f}{R-T}$, we observe that the optimal solution is $\tilde{\mu}_v^{BV}$ without constraints (12)–(14) (it directly follows from Lemma B.1). Next, we have to regularize μ_v^{BV*} to make (12)–(14) valid. Recall the constraints (12)–(14):

$$\frac{\theta_v}{\mu_v}+\frac{\delta\theta_f}{\mu_f} \le R-\delta T, \quad \mu_f-\delta\mu_v \le \frac{\theta_f}{R-T}-\frac{\delta\theta_v}{(1-\delta)R},$$

$$\frac{\theta_v}{\mu_v - \lambda} + \frac{\delta\theta_f}{\mu_f - \delta\lambda} \geq R - \delta T \text{ for } \mu_v > \lambda \text{ and } \mu_f > \delta\lambda.$$

The first condition must hold (as the solution makes it binding), thus no revenue is generated. If the second constraint is not satisfied, the boundary solution would be the optimal one: $\frac{1}{1+\delta}(\mu - \frac{\theta_f}{R-T} + \frac{\delta\theta_v}{(1-\delta)R})$. If the third constraint is not satisfied, Region V dominates Region BV. In this case, we require that $\lambda_v^{BV*} \leq \lambda$. Q.E.D.

Proof of Lemma B.1: Recall that given (μ_f, μ_v) which satisfies (B.1) and (12)–(14), we have

$$\lambda_v^{BV} = \frac{\mathcal{B} - \sqrt{\mathcal{D}^2 + 4\delta^2\theta_f\theta_v}}{2\mathcal{A}}.$$

By replacing μ_f with $\mu - \mu_v$, we have

$$\frac{\partial\mathcal{B}}{\partial\mu_v} = -(R - \delta T)(1-\delta) < 0, \quad \text{and} \quad \frac{\partial\mathcal{D}}{\partial\mu_v} = -(R - \delta T)(1+\delta) < 0.$$

Therefore, when $\mathcal{D} < 0$, λ_v^{BV} is decreasing in μ_v. Note that $\mathcal{A}$ is independent of μ_v.

When $\mathcal{D} \geq 0$,

$$\frac{\partial\lambda_v^{BV}}{\partial\mu_v} = -\frac{-(1-\delta)}{2\delta} + \frac{(1+\delta)\mathcal{D}}{2\delta\sqrt{\mathcal{D}^2 + 4\delta^2\theta_f\theta_v}}.$$

Since $\mathcal{D} \geq 0$, then $\frac{\partial\lambda_v^{BV}}{\partial\mu_v}$ is increasing in $\mathcal{D}$. And when $\mathcal{D} = 0$, $\frac{\partial\lambda_v^{BV}}{\partial\mu_v} = -\frac{-(1-\delta)}{2\delta} < 0$. Let $\bar{D}$ solves $\frac{\partial\lambda_v^{BV}}{\partial\mu_v} = 0$, we must have $\bar{D} > 0$. So when $0 \leq \mathcal{D} < \bar{D}$, $\frac{\partial\lambda_v^{BV}}{\partial\mu_v} < 0$; and when $\mathcal{D} > \bar{D}$, $\frac{\partial\lambda_v^{BV}}{\partial\mu_v} > 0$.

When $\mu_v = \tilde{\mu}_v^{BV} = \frac{\mu}{1+\delta} - \frac{\delta(\theta_f - \theta_v) + (1-\delta)\sqrt{\delta\theta_f\theta_v}}{(1+\delta)(R-\delta T)}$, $\mathcal{D} = \bar{D}$. Since $\mathcal{D}$ is decreasing in μ_v, we can conclude that when $\mu_v \leq \tilde{\mu}_v^{BV}$, λ_v^{BV} is increasing in μ_v; when $\mu_v \geq \tilde{\mu}_v^{BV}$, λ_v^{BV} is decreasing in μ_v. Q.E.D.

Proof of Proposition 3: Before characterizing the optimal capacity allocation in Region VF, we first provide an auxiliary result. Lemma B.2 proves that the unique effective arrival rate to the virtual channel (which is defined in (B.4) and here we denote it as λ_v^{VF}) is increasing in the capacity allocated to this channel.

Lemma B.2. *Given (μ_f, μ_v) which satisfies (B.3) and (15)–(18), λ_v^{VF} is increasing in μ_v.*

Since λ_v^{VF} is increasing in μ_v and the objective function is monotone in λ_v^{VF}, the optimal solution must be in the boundaries if $r_v \neq r_f$. Q.E.D.

Proof of Lemma B.2: Recall that given (μ_f, μ_v) which satisfies (B.3) and (15)–(18), we have

$$\lambda_v^{VF} = \frac{-\tilde{\mathcal{B}} + \sqrt{\tilde{\mathcal{D}}^2 + 4\theta_f\theta_v}}{2\tilde{\mathcal{A}}}.$$

By replacing μ_f with $\mu - \mu_v$, we have

$$\frac{\partial\tilde{\mathcal{B}}}{\partial\mu_v} = -(2-\delta)T < 0, \quad \text{and} \quad \frac{\partial\tilde{\mathcal{D}}}{\partial\mu_v} = -\delta T < 0.$$

Then,

$$\frac{\partial(-\tilde{\mathcal{B}} + \sqrt{\tilde{\mathcal{D}}^2 + 4\theta_f\theta_v})}{\partial\mu_v} = (2-\delta)T - \frac{\delta T\tilde{\mathcal{D}}}{\sqrt{\mathcal{D}^2 + 4\theta_f\theta_v}},$$

which is decreasing in $\tilde{\mathcal{D}}$. When $\tilde{\mathcal{D}} \to \infty$, $\frac{\partial(-\tilde{\mathcal{B}} + \sqrt{\tilde{\mathcal{D}}^2 + 4\theta_f\theta_v})}{\partial\mu_v} \to 2(1-\delta)T > 0$. So λ_v^{VF} is increasing in μ_v. Q.E.D.

Proof of Theorem 2: As illustrated in Figure 3, each system size includes several equilibrium regions. The optimal capacity allocation for each size is the one that yields the greatest profit among the regions it includes:

1. **Extreme Small systems** include Regions B and BV if $\frac{\theta_f}{R-T} \le \underline{\mu}^{BV}$, thus the optimal capacity allocation must be in Region BV; and if $\frac{\theta_f}{R-T} > \underline{\mu}^{BV}$, it includes Regions B and BF, thus the optimal capacity allocation must be in Region BF.

2. **Small systems** include Regions B, BV, BVF, BF and F. Region B is dominated by Regions BV and BF. From Proposition 1 we know that the optimal solution in Region BVF is on the boundaries (in this case, either on the boundary with BV or on the boundary with BF). When Region F is applicable ($\lambda \le \mu - \frac{\theta_f}{R-T}$), it dominates Region BF. The optimal value function is $r_f\lambda$ in Region F, and $r_f\left(\mu - \frac{\theta_f}{R-T}\right)$ in Region BF. In both cases the optimal allocation is $\mu_f = \mu$ and $\mu_v = 0$. The optimal value function in Regions F and BF is, therefore, $r_f\left(\min\{\lambda, \mu - \frac{\theta_f}{R-T}\}\right)$. Comparing the latter with the optimal value function in Region BV yields the condition as stated.

3. **Medium systems** include Regions B, BV, BVF, VF and F. Region B is dominated by Region BV. From Proposition 1 we know that the optimal solution in Region BVF is dominated by either BV or VF. Moreover, Region F is dominated by VF (as an extreme case in VF). Therefore, it suffices to compare the optimal value function in Regions BV and VF, as stated.

4. **Large systems** include Regions B, BV, V, VF and F. Region B is dominated by Region BV and Regions BV and VF are dominated by Region V (as an extreme case in VF). Therefore, the optimal region is VF, as stated. Since the solution of Region VF is a boundary one, it will utilize either the face-to-face channel or the virtual channel while supporting returning patients.

Q.E.D.

Proof of Corollary 1: The results follow Theorem 1 if we replace T by $T - b$. Q.E.D.

Proof of Corollary 2: First let us replace T by $T - b$ in the results of Theorem 2. Then, the optimal capacity allocation must lead to the six scenarios presented in the Corollary statement. The results in each scenario are straightforward. It is worth noting that $\frac{\delta(\theta_f-\theta_v)+(1-\delta)\sqrt{\delta\theta_f\theta_v}}{R-\delta T+\delta b} \le \frac{\theta_f}{R-T+b} - \frac{\delta\theta_v}{(1-\delta)R}$ is equivalent to the second condition in Proposition 2 (i.e., $(\mu - \tilde{\mu}_v^{BV}, \tilde{\mu}_v^{BV}) \in$ Region BV). Q.E.D.

Proof of Lemma 1: When $\mu \ge (1+\delta)\lambda + \frac{\theta_v}{R(1-\delta)} + \frac{\theta_f}{R-T+b}$, the system performs as large one. $\bar{b}$ is the one to make this condition hold. Q.E.D.

Proof of Proposition 4:

- **The *PURE* Scenario**: It is straightforward that setting no bonus is optimal.
- **Scenario *NV***: It can be verified that $(r_f - b)\left(\mu - \frac{\theta_f}{R-T+b}\right)$ is a concave function of b and the first order condition lead to the solution $\sqrt{\frac{\theta_f(R-T+r_f)}{\mu}} - R + T$. It is worth noting that $\bar{b}_{NV}$ is an upper bound for the optimal b. Then, it is straightforward that b^{NV*} is limited by its upper and lower bounds.

- **Scenario $\boldsymbol{NF}$**: The results are straightforward.
- **Scenario $\boldsymbol{NB}$**: The first order derivative of the objective function is

$$(1-\delta)(\mu-\lambda)-\delta\lambda-\frac{\theta_v}{R}+\frac{\theta_f(r_v-r_f-(1-\delta)(R-T))}{(R-T+b)^2}.$$

—If $r_v-r_f-(1-\delta)(R-T)\leq 0$, the objective function is convex, so the optimal solution must be on the boundary (the first case of b^{NB*}).

—If $r_v-r_f-(1-\delta)(R-T)>0$ and $(1-\delta)(\mu-\lambda)-\delta\lambda-\frac{\theta_v}{R}\geq 0$, the objective function is increasing in b, so the optimal solution is to set b as large as possible (the first case of $b^{NB\#}$).

—If $r_v-r_f-(1-\delta)(R-T)>0$ and $(1-\delta)(\mu-\lambda)-\delta\lambda-\frac{\theta_v}{R}<0$, the second order derivative is negative; the first order condition then leads to the second case of $b^{NB\#}$. The second case for b^{NB*} is restricted by its upper and lower bounds. Q.E.D.

Proof of Proposition 5: Follows directly from the definition of the thresholds for μ. Q.E.D.

Proof of Corollary 3: When $r_f\geq r_f$ and $\mu\geq\lambda+\frac{\theta_f}{R-T}$, the system is able to achieve the highest reimbursement $r_f\lambda$. Q.E.D.

Proof of Proposition 6:

1. **When $\boldsymbol{r_v\leq r_f}$**: If $\mu\geq\lambda+\theta_f/(R-T)$, the $PURE-F$ Scenario generates the largest reimbursement $r_f\lambda$ with $b^*=0$. We, therefore, need to consider only the case where $\mu<\lambda+\theta_f/(R-T)$. In this Case, the reimbursement in Scenario NV is $r_f(\mu-\frac{\theta_f}{R-T})$, and $r_v\left(\frac{\mu}{1+\delta}-\frac{\left(\sqrt{\delta\theta_f}+\sqrt{\theta_v}\right)^2}{(1+\delta)(R-\delta T)}\right)$ is the upper bound of the reimbursement in Scenario NF. Since (30) holds, Scenario NF is dominated by Scenario NV without a bonus. Moreover, since $r_v\leq r_f$, Scenario NB is dominated by Scenario NV. So without a bonus, the optimal solution must lead to Scenario NV, $PURE-B$, or $PURE-V$. The only possible case for the access rate to decrease is when the optimal solution appears in Scenario NV without a bonus whereas Scenario NF is the optimal solution with a bonus. In this case, however, since $r_v\leq r_f$, the resulting λ_v^* must be greater than the original access rate without a bonus.

2. **When $\boldsymbol{r_v\geq(1+\delta)r_f}$**: Per Proposition 3, Scenario NB is dominated by Scenario NF for any b. Note that $r_f\left(\mu-\theta_f/(R-T)\right)$ is the upper bound for the revenue in Scenario NV, and $r_v\left(\frac{\mu}{1+\delta}-\frac{\theta_f}{(1+\delta)(R-T)}-\frac{\theta_v}{(1+\delta)(1-\delta)R}\right)$ is the lower bound for the revenue in Scenario NF. When condition (31) holds, Scenario NV is dominated by Scenario NF without bonus. In this case, the optimal solution must lead to Scenarios NF, $PURE-B$, or $PURE-V$. Clearly, it is possible to have a decreased access rate when the optimal solution is in Scenario NF without a bonus, but it is in Scenario NV with a bonus. In this case, since $r_f\leq r_v$, the resulted λ_f^* must be greater than the original access rate. Q.E.D.

Proof of Proposition A.1: Note that the sign of

$$\frac{\delta(\theta_f-\theta_v)+(1-\delta)\sqrt{\delta\theta_f\theta_v}}{(R-\delta T+\delta b)}\frac{\theta_f}{R-T+b}+\frac{\delta\theta_v}{(1-\delta)R}$$

is as same as the sign of

$$\left(\delta(\theta_f-\theta_v)+(1-\delta)\sqrt{\delta\theta_f\theta_v}\right)(R-T+b)-\theta_f(R-\delta T+\delta b)+\frac{\delta\theta_v(R-\delta T+\delta b)(R-T+b)}{(1-\delta)R}, \tag{B.6}$$

as $R-\delta T>0$ and $R-T+b>0$. It can be verified that (B.6) is a quadratic convex function of b. Therefore, if Assumption A.1 holds, i.e.,

$$\frac{\delta(\theta_f-\theta_v)+(1-\delta)\sqrt{\delta\theta_f\theta_v}}{(R-\delta T)}<\frac{\theta_f}{R-T}-\frac{\delta\theta_v}{(1-\delta)R},$$

then there exists a unique $b^0\geq 0$, such that (B.6) equals to 0 when $b=b^0$. In other words, under Assumption A.1, there exists a unique $b^0\geq 0$, such that

$$\frac{\delta(\theta_f-\theta_v)+(1-\delta)\sqrt{\delta\theta_f\theta_v}}{R-\delta T+\delta b^0}=\frac{\theta_f}{R-T+b^0}-\frac{\delta\theta_v}{(1-\delta)R}.$$

Thus, when $b\leq b^0$, the optimal capacity allocation results in the first case of (21); when $b>b^0$, however, the optimal capacity allocation results in the second case of (21). We therefore have

$$\lambda_v^{NF}(b)=\begin{cases}\min\left\{\lambda,\, \frac{\mu}{1+\delta}-\frac{(\sqrt{\delta\theta_f}+\sqrt{\theta_v})^2}{(1+\delta)(R-\delta T+\delta b)}\right\}, & \text{if } b\leq b^0,\\ \min\left\{\lambda,\, \frac{\mu}{1+\delta}-\frac{\theta_f}{(1+\delta)(R-T+b)}-\frac{\theta_v}{(1+\delta)(1-\delta)R}\right\}, & \text{otherwise.}\end{cases} \tag{B.7}$$

Note that both

$$(r_v-\delta b)(\frac{\mu}{1+\delta}-\frac{(\sqrt{\delta\theta_f}+\sqrt{\theta_v})^2}{(1+\delta)(R-\delta T+\delta b)}), \tag{B.8}$$

and

$$(r_v-\delta b)\frac{\mu}{1+\delta}-\frac{\theta_f}{(1+\delta)(R-T+b)}-\frac{\theta_v}{(1+\delta)(1-\delta)R} \tag{B.9}$$

are concave.

Furthermore, when $[(b_{BV})^+,\ \min\{r_v/\delta,\ \bar{b}\}]\neq\emptyset$,

$$b^{NF*}=\min\left\{\max\left\{b^{NF\#},\ (b_{NF})^+\right\},\ r_v/\delta,\ \bar{b}\right\},$$

where

$$b^{NF\#}=\begin{cases}b^1_{NF}, & \text{if } b^1_{NF}\leq b^0\\ b^2_{NF}, & \text{otherwise.}\end{cases}$$

Here, b^1_{NF} and b^2_{NF} are the optimal solutions to the two cases of $(r_v-\delta b)\lambda_v^{NF}(b)$. Specifically,

$$b^1_{NF}=\min\left\{\frac{(\sqrt{\delta\theta_f}+\sqrt{\theta_v})\sqrt{R-\delta T+r_v}}{\delta\sqrt{\mu}}-\frac{R-\delta T}{\delta},\ \bar{b}^1_{NF}\right\},$$

where the first term in the minimum comes from taking the first order condition of (B.8), and

$$\bar{b}^1_{NF}=\begin{cases}\infty, & \text{if } \lambda\geq\frac{\mu}{1+\delta}\\ \frac{\left(\sqrt{\delta\theta_f}+\sqrt{\theta_v}\right)^2}{\delta(\mu-(1+\delta)\lambda)}-\frac{R-\delta T}{\delta}, & \text{otherwise;}\end{cases}$$

$$b^2_{NF}=\min\left\{\sqrt{\frac{\theta_f(\delta R-\delta T+r_v)}{\delta\mu-\frac{\delta\theta_v}{(1-\delta)R}}}-R+T,\ \bar{b}^2_{NF}\right\},$$

where the first term in the minimum comes from taking the first order condition for (B.9), and

$$\bar{b}^2_{NF} = \begin{cases} \infty, & \text{if } \lambda \geq \frac{\mu}{1+\delta} - \frac{\theta_v}{(1+\delta)(1-\delta)R} \\ \frac{\theta_f}{\mu-(1+\delta)\lambda-\frac{\theta_v}{(1-\delta)R}} - R + T, & \text{otherwise.} \end{cases}$$

We conclude the proof by showing that b^1_{NF} generates the highest revenue when $b^1_{NF} \leq b^0$, while otherwise, b^2_{NF} generates the highest revenue; namely,

$$\min\left\{\lambda, \frac{\mu}{1+\delta} - \frac{(\sqrt{\delta\theta_f}+\sqrt{\theta_v})^2}{(1+\delta)(R-\delta T+\delta b)}\right\} \geq \min\left\{\lambda, \frac{\mu}{1+\delta} - \frac{\theta_f}{(1+\delta)(R-T+b)} - \frac{\theta_v}{(1+\delta)(1-\delta)R}\right\}.$$

When $b = b^0$, the inequality is binding. Q.E.D.

Appendix References

Shi, Z., A. Mehrotra, C.A. Gidengil, S.J. Poon, L. Uscher-Pines, K.N Ray. 2018. Quality of care for acute respiratory infections during direct-to-consumer telemedicine visits for adults. *Health Affairs* **37**(12) 2014–2023.

Uscher-Pines, L., R. Malsberger, L. Burgette, A. Mulcahy, A. Mehrotra. 2016. Effect of teledermatology on access to dermatology care among medicaid enrollees. *JAMA dermatology* **152**(8) 905–912.

Yamamoto, D.H. 2014. Assessment of the feasibility and cost of replacing in-person care with acute care telehealth services. *Alliance for Connected Care, December* .