Service Systems with Heterogeneous Customers: Investigating the Effect of Telemedicine on Chronic Care

Balaraman Rajan
California State University East Bay
balaraman.rajan@csueastbay.edu

Tolga Tezcan
London Business School
ttezcan@london.edu

Abraham Seidmann
University of Rochester, Simon School of Business
avi.seidmann@simon.rochester.edu

Medical specialists treating chronic conditions typically face a heterogeneous set of patients. Such heterogeneity arises because of differences in medical conditions as well as the travel burden each patient faces to visit the clinic periodically. Given this heterogeneity, we compare the strategic behavior of revenue-maximizing and welfare-maximizing specialists and prove that the former will serve a smaller patient population, spend more time with the patients, and have shorter waiting times. We also analyze the impact of telemedicine technology on patient utility and the specialists' operating decisions. We consider both the case when specialists can freely set their own fee for service and the case when fees are set exogenously by a third-party payer. We prove that with the introduction of telemedicine the specialists become more productive and the overall social welfare increases, though some patients, unexpectedly, will be worse off. Our analytical results lead to some important policy implications for facilitating the further deployment of telemedicine in the care of chronically ill patients.

1. Introduction

Service quality, unlike product quality, largely depends on perception. Service providers have employed different mechanisms to signal quality, ranging from price (**?**) to service environment or ambiance (**?**). The nature of the server-customer interaction adds another vital dimension. In a healthcare setting, the interaction between the service provider (such as a medical specialist) and the customer (patient), or more technically the consultation time, gains further importance. Most patients form opinions based on their interaction with the specialist since they are poorly equipped to judge the long-run clinical outcomes of any particular visit. Even though a specialist might be able to cover all possible clinical aspects in a short consultation, if the specialist spends very little

time with patients, the patients might not be satisfied with the specialist's effort and may feel they are being rushed out (**??**).

Medical specialists, like many other service providers in almost all settings, face an inherently challenging trade-off: While a customer would like more time with the provider, spending more time with each customer reduces the number of customers who can be served in a given day and reduces the number of billed visits per day, and the slower service rate delays other customers waiting to be served. Longer waiting times reduce customer satisfaction (**?**), and more so if the waiting time is significant when compared to the actual service time. Yet it is common in medical settings that patients have to wait for hours for only a few minutes of interaction with their provider. In this paper we study the speed-quality trade-off when chronically ill patients have heterogeneous treatment utility. Considering that specialists are also busy and have a huge backlog of appointments, we also explore why and when a busy specialist would adopt telemedicine to begin with and analyze the impact of telemedicine on this trade-off in that setting.

Features of chronic care: This paper is motivated by our long-term study of the clinical service requirements of chronically ill patients. Chronic conditions represent a huge burden on the modern healthcare system, and they are a major factor in mortality and disability (**?**). About 140 million Americans have at least one chronic condition, and these patients account for about 75% of total U.S. health expenditure (**?**). These patients require lifelong care by highly trained specialists and must visit their providers periodically to adjust their medication and dosages and to manage their conditions. For example, a patient suffering from Parkinson's disease (PD) might visit their specialist two to six times a year and visits are sometimes added or missed due to unexpected disease progression, severe complications, or other nontrivial comorbidities (see Appendix **??** that describes care of a typical patient suffering from this condition).

Chronically ill patients are not homogeneous in the utility they derive from treatment by specialists, which is one of the main motivations of the models we study in this paper. Most specialists are concentrated in urban areas, and there are few or even none in remote or rural areas (**?**). Thus patients located in urban areas are closer to specialists, but many rural patients travel long distances to visit one. In fact, a one-way travel time of 4 hours is not uncommon for a visit of just 30 minutes for patients from rural areas (**?**). Besides distance, there are also other reasons for heterogeneity among patients, such as their overall morbidity. Our framework in this paper accommodates a large number of patients and specialists, and it also fits well with major chronic conditions such as diabetes, rheumatoid and psoriatic arthritis, and kidney disease.

Recently, telemedicine has proved to be a feasible alternative to office visits to treat chronically ill patients. Consultations related to chronic conditions are required multiple times a year to manage the underlying disease, and the medical condition of a patient deteriorates without these periodic visits. For many medical conditions,[1] it is clinically feasible to substitute virtual visits via a telemedicine facility for some in-person visits to a medical practitioner. Specifically in chronic care, telemedicine is used for direct, synchronous, and remote communication between the physician and the patient.[2] Therefore, the introduction of telemedicine has the potential to increase patient utility by reducing the travel burden and in the process fundamentally change the operational challenges specialists face.

Impact of treatment value heterogeneity on specialists' actions: Motivated by the heterogeneity in treatment utility, the first part of this paper establishes the impact of this heterogeneity on patient and specialist behaviors. Service systems in which rational customers decide whether to seek service based on service quality have been studied extensively (**????**). **?** recently considered the speed-quality trade-off for homogeneous customers. However, their results do not cover the case when the patients are heterogeneous, which is precisely the case with the treatment of chronic patients.

In the first part of the paper, we prove that if the patient population is heterogeneous, the operating policies of revenue-maximizing specialists and welfare-maximizing specialists are such that the former work slower and treat fewer patients than the latter. We also show that the utilization of the welfare-maximizing specialists is higher than that of the revenue-maximizing specialists, resulting in relatively longer wait times for the patients and a busier appointment calendar. We also find that as the travel burden increases for the patients, specialists tend to compensate by spending more time with them, even though lowering the service rate might increase congestion and reduce the number of patients seen per day, to optimize their overall revenue.

Impact of telemedicine on specialists' and patients' actions: In the second part of the paper we study the impact of telemedicine on the quality-speed trade-off in chronic care. We extend our basic model to include the case in which patients have the option to seek treatment by their specialist via telemedicine. We seek answers to the following fundamental questions regarding the

[1] Telemedicine has already been used for pediatrics, psychiatry, movement disorders, neurological disorders like Alzheimer's disease and epilepsy, dermatological disorders, and such chronic disorders as diabetes; see **?** and **?** for various applications, and **?** for a review of telemedicine's benefits.

[2] Telemedicine can take a number of different forms. There can be asynchronous information exchange, such as when the provider is remote from the facility, as with medical diagnostics and radiology. Telemedicine can support peer-to-peer consultation (for instance, via a video-conferencing facility, as with telestroke), or it can involve direct synchronous communication between the physician and the patient (**?**).

operational issues: Given the advantages and limitations of telemedicine, what clinical fields are likely to benefit most? Why should an already busy specialist add a telemedicine service? Also, providers can react to the introduction of telemedicine by changing their prices and service rates. What will be the economic impact of these changes on providers and patients? More specifically, given the severe shortage of specialists even at the current level of coverage (**?**), how will an increase in demand for specialist care because of telemedicine be handled? Will specialists choose to accept more patients or try to leverage the increase in expected utility (surplus) for each patient by charging higher fees? How will service levels (utilization) change? Finally, what will be the subsequent process impact on patients and on social welfare?

By incorporating a factor for clinical feasibility, we find that with the introduction of telemedicine, the utilization of the revenue-maximizing specialists goes up, their service rate increases, their overall productivity in terms of clinical capacity increases, and (if the perceived quality of the telemedicine visits is comparable to that of the in-person visits) the average price per visit decreases.

In addition, we establish a relatively simple sufficient condition to identify when telemedicine becomes economically feasible. There are three important parameters that determine this feasibility: the perceived "quality" difference between telemedicine and in-person visits, the travel burden (or disutility) because of the distance the patients have to travel to the clinic (without telemedicine), and the technological and maintenance costs for a patient to receive telemedicine treatment. Interestingly enough, the same condition is sufficient for the introduction of telemedicine to increase the total welfare. However, we find that the benefit from telemedicine is not uniform for all patients, and some patients who continue to use in-person visits may be worse off.

We then make general observations as to which characteristics are necessary for telemedicine to be attractive for a certain specialty. For example, based on our ongoing research experience with Parkinson's patients, patients could expect to receive quality care with telemedicine, at least for a certain proportion of their annual visits (**?**). Also, many Parkinson's patients have a relatively high travel cost, because of their motion disorder conditions, and leading neurologists are usually concentrated in large urban areas (**?**). Therefore, treatment of Parkinson's patients is highly suitable for implementing telemedicine, assuming the technological cost is not excessive.

Technical contributions: Aside from our findings on the impact of patient heterogeneity and telemedicine on the speed-quality trade-off, this paper makes two main technical contributions. First, we extend the literature on the analysis of queuing systems with rational customers. Due to customer heterogeneity, the standard results from the literature cannot be used directly. In

our proof we first show that the revenue-maximizing service provider's objective can be expressed in a simpler form. Then we use the implicit function theorem to establish the optimal decisions for revenue-maximizing and welfare-maximizing service providers. Our second contribution is to establish the equilibrium behavior of patients when telemedicine is introduced. For these more complex analyses we use the concept of non-atomic games (reviewed in Appendix **??**). To the best of our knowledge our study is the first paper in the queuing literature that uses this concept. We have chosen this approach because non-atomic games enable us to model and analyze a conflict situation where no single patient has an influence on the final outcome, but the aggregate behavior of a "large" set of patients would change the payoffs. Using this approach, we establish the unique Nash equilibrium in our queuing model and then use this result to find the optimal operating decisions for the specialist who will be running a hybrid modality: treating some patients face-to-face and others via a mix of telemedicine and face-to-face visits at the clinic. This approach also allows us to analytically characterize the impact of telemedicine on patient access to care, patient wait and visit times, and physician utilization, capacity, fees, and revenues, as well as the impact of telemedicine on total welfare.

2. Related Literature

Our research is closely related to the stream of literature on queues with rational customers. **?** was the first to demonstrate that in a single-server Markovian queuing system, when customers are rational and decide to seek service based on their utility, they end up joining at more than the socially optimal levels. He also showed that this behavior can be curbed by charging customers a toll. **?** extended this result to unobservable queues. Later, **?** and **?** extended the results in **?** and **?**, respectively, to the case with heterogeneous customers.

In most service systems, the service provider can also alter the service speed to manage congestion. **?** considered optimizing over toll price and capacity (service rate), with the cost of capacity being convex in the service rate and **?** consider a similar model with linear capacity costs. They show how the revenue-maximizing server's actions deviate from those of a welfare-maximizing server, in a similar spirit to **?** and **?**. However, because **?** focused on computer systems, he did not model the impact of service speed on customer's utility.

The pioneering paper by **?** modeled the impact of service speed explicitly and examined the interaction between service value and speed. They considered a single server serving homogeneous customers and studied the quality effect of a given service rate and incorporated it into the customer's decision criteria. In their model, the probability that a customer decides to seek treatment

is based on the value of the service at the given service rate (the quality effect), the waiting cost incurred, and the price to be paid. They show that the quality effect may change the server's decisions significantly. However, **?** did not consider customers with heterogeneous service valuation, and in the healthcare setting patients typically have to travel from different locations to see a provider. Differences in distances naturally cause heterogeneous service values, and we extend the literature by incorporating this into a model similar to that in **?** and **?**.

A service provider can offer different service levels (with different waiting times) and different prices to obtain more revenue from heterogeneous customers. **?** derive a pricing mechanism that is optimal and incentive-compatible for a welfare-maximizing provider, and **?** (see the references therein for an extensive review of this literature) derives a pricing mechanism for a revenue-maximizing provider. Although these studies modeled customer heterogeneity explicitly, they did not study the case when the customer's service valuation depends on service speed as well. In addition, in our model the server can offer two modes of service, in-person and telemedicine, and customers prefer more time with the server. Also, customers' preference for a particular type of service mode might be different based on their distance, their, whereas in **?** and **?** all customers prefer waiting less. We refer the reader to **?** and **?** for extensive reviews of the literature on queues with rational customers.

We also contribute to the literature on the economic feasibility of telemedicine. In addition to evaluating the clinical feasibility of telemedicine (see several works cited in our introduction), the cost effectiveness and socio-economic benefits of telemedicine have been investigated empirically in various studies; see review papers by **?** and **?**. The impact of telemedicine in primary care has also been the subject of study in the current literature (**??**). The general conclusion is that although there is evidence in certain medical fields that telemedicine is beneficial, the economic impact of telemedicine is unclear. We hope to clarify this impact with a service model that considers the main trade-offs that arise with the introduction of telemedicine.

3. Effect of patient heterogeneity on specialist interaction

In this section, we analyze the impact of patient heterogeneity on patient and provider actions. We first consider a basic model without telemedicine, before turning in Section **??** to the effect of telemedicine on patient and provider behavior. After introducing our analytical model, we establish the properties of the equilibrium behavior of rational customers and revenue- and welfare-maximizing specialists. We then establish the impact of increasing traveling cost on a revenue-maximizing specialist's optimal actions. Then we compare the optimal actions of revenue- and welfare-maximizing specialists. Finally, we analyze the case when prices are determined exogenously.

3.1. Patient utility and the specialist's objective

We consider a single specialist (monopolist) in a region serving a patient base. The specialist chooses his *service rate* and *price* in order to either maximize his revenue (see Section **??**) or total welfare (see Section **??**). In response to the specialist's choices, patients decide to seek service based on their expected net utility. Each patient's utility comprises (i) a reward from seeking treatment, (ii) a quality cost, (iii) a payment, and (iv) a congestion cost. We next describe each component of the customer's utility function, assuming the service rate and price are fixed, before we describe the details of the specialist's objective function (see Appendix **??** for a detailed description of a Parkinson patient's profile).

Reward from seeking treatment: We model the patient's utility from seeking treatment as a function of the patient's distance from the specialist. This is because, typically, a specialist has patients coming from various places and the patient's net utility from seeking service depends on the distance to be traveled to see the specialist. We are also interested in this setting due to our ultimate goal of understanding the effect of telemedicine on patient and provider choices (see Section **??**), and one of the most significant benefits of telemedicine is that it obviates the need to travel for some consultations with the specialist.

In order to model the impact of distance on the patient's utility, we assume that the potential patients are indexed by x, their distance from the specialist. Also, we assume that each patient gets a constant reward m per visit to the specialist. Specifically, let $t : \mathbb{R}_+ \to \mathbb{R}_+$, where $t(x)$ denotes the travel burden for a patient located at a distance x from the specialist.[3] We assume that the net benefit per visit to the specialist, after accounting for the travel burden, for a patient at a distance x is given by $m - t(x)$. Since all patients see the same specialist and suffer from a similar medical condition, for simplicity, we set aside differences in utility owing to differences in current health status or individual perceptions. (This assumption can be relaxed easily; see Remark **??** below.)

We make the following assumptions in our model. We assume that t is strictly increasing and thus invertible, and (without loss of generality) that $t(0) \geq 0$. We let f denote the density function for the distribution of the distance of the patients from the specialist and F denote the associated cdf. Throughout, we assume that F and t satisfy the conditions in Appendix **??**, unless stated otherwise. We set $M_v = m - t(0)$ and assume that $M_v > 0$; that is, at least some of the patients would potentially seek treatment. Also, we denote by X_m the farthest distance of the patient whose benefit from treatment net of travel burden is zero; hence $m - t(X_m) = 0$. We note that patients whose distance is more than X_m would never seek treatment.

[3] We use $\mathbb{R}$ to denote the real numbers, and $\mathbb{R}_+ = [0, \infty)$.

Unlike in the standard queuing models, each patient needs repeated visits to the specialist in chronic care. We assume implicitly that all patients need to see the specialist at the same rate on average. This should hold true for a population of patients suffering from the same chronic disease. In addition, this assumption can be relaxed, and the arrival rate can be taken as the average rate of patient visits at the specialist, if the visit rate is not correlated with the distance or the treatment utility that patients receive. We denote by $\Lambda < \infty$ the total number of patients as well as the total arrival rate if the specialist decides to serve all the patients.

Quality cost: We use $Q : \mathbb{R}_+ \to \mathbb{R}_+$ to denote the quality cost[4] or disutility to the patient as a function of the service rate $\mu(\geq 0)$. We assume without loss of generality that $Q \geq 0$ is increasing (see Appendix **??** for other technical assumptions on Q). In other words, patients prefer more time with the specialist (**?**).

Payment: We assume that patients incur a monetary cost every time they visit a provider. Since patients typically are covered by insurance, most of the cost is borne by their insurance. Thus, if the price per visit is p for the services rendered, the patient pays βp, for $0 < \beta \leq 1$, where β is the co-insurance rate. Any co-payments or other out-of-pocket expenses are absorbed into the heterogeneous benefit $m - t(x)$ to keep our modeling concise.

Congestion cost: We use the term $c\mathbb{E}W(\mu, \lambda)$ to capture the congestion cost, where μ is the service rate chosen by the specialist and λ is the equilibrium arrival rate, whose specifics are discussed below. This term measures the disutility of patients who are dealing with a busier specialist, where $\mathbb{E}W(\mu, \lambda)$ is the expected delay and c is the opportunity cost per unit time of "delay" incurred by the patient. This cost should be viewed as a notional cost in general. Some patients may need to see the doctor sooner than the recommended time, and some patients may not comply with the recommendations (they miss scheduling their next visit, which might trigger a complication that may require one or more unexpected and unscheduled visits to the clinic). Thus, the patient appointments could be regular appointments (which, as we have noticed, need not be precisely at the recommended time, as there is some uncertainty with respect to the specialist's availability or the patient's own medical condition), or it could be for emergency appointments or new appointments in case of cancellations (see Appendix **??** for an example of a patient's appointment log). The term thus captures the fact that the patient's utility decreases as the provider gets busier and

[4] Utility functions are generally measured using standard methods described in the consumer research literature (see, for example, (**??**). Customers are typically asked for alternatives they would choose (for example, in this case, 10 minutes of extra time with their specialist or a 45-minute physiotherapy session). If one alternative could be assigned a monetary value (a 45-minute physiotherapy session could be worth, say, $150), then the patient's choice in each such pair could be used to measure the utility function. The utility could also be measured by observing customer choices when they are offered different service times along with differential co-payments.

busier—longer waiting times for visit rescheduling, medication renewals or adjustments, physical therapy referrals, or other unplanned visit requests in the case of any unforeseen concern.

The term $\mathbb{E}W(\mu,\lambda)$ thus enables us to incorporate the important operational constraint that at steady state, specialists cannot plan to operate at a very high utilization. This follows from several notable works in the literature on the optimal panel size. **?** note that "the primary lever to bring demand and supply into a relationship that is compatible with being able to offer short appointment dates is patient panel size." As a result, we noticed that specialists who deal with chronic patients operate with a relatively small panel size as they try to maintain acceptable utilization and service levels. For example, one of the specialists we worked with limits his panel size to only 200 active patients as compared to 2,000-2,500 patients for a primary care physician (see Appendix **??** for the description of a sample clinic). With the current geographical distribution and small panel sizes, it is no wonder that 42% of Medicare beneficiaries diagnosed with Parkinson's do not even see a general neurologist, not to mention a specialist (**?**).

Patient's utility function: Given that the specialist chooses a service rate μ and price p, the arrival rate, λ, is obtained from the number of patients who decide to seek treatment based on their utility in equilibrium, whose specifics are discussed in the next section. The total utility, $\Psi(:[0,\Lambda]\times\mathbb{R}\times\mathbb{R}_+\times[0,M_v]\to\mathbb{R})$, for a patient at distance x as a function of equilibrium arrival rate λ, price p, and service rate μ is given by

$$\Psi(\lambda,p,\mu,x) = \underbrace{m-t(x)}_{\substack{\text{reward from}\\ \text{seeking treatment}}} - \underbrace{Q(\mu)}_{\substack{\text{quality}\\ \text{cost}}} - \underbrace{\beta p}_{\text{payment}} - \underbrace{c\mathbb{E}W(\mu,\lambda)}_{\substack{\text{congestion}\\ \text{cost}}}. \tag{1}$$

We model the demand from patients as a Poisson process; that is, the time between arrivals for patient appointments has an exponential distribution. We assume that the service times offered by the single specialist are exponentially distributed as well (see **?** and **?** for similar assumptions), resulting in an M/M/1 queuing model, an assumption that helps provide analytical tractability. With the M/M/1 assumption, $\mathbb{E}W(\mu,\lambda)=1/(\mu-\lambda)$. Our results can readily be extended to cover other interarrival and service-time distributions using standard queuing approximations, and we present other extensions to this model (including considering just the waiting time in queue, $(1/(\mu-\lambda)-1/\mu)$ in Appendix **??**. We also assume that patients do not renege; that is, our model does not allow no-shows.

Remark 1. Our utility model builds on extant literature. While **?** and **?** consider customer heterogeneity, they do not consider the quality cost of service, $Q(\mu)$. **?** consider $Q(\mu)$ to be linear but assume the benefit to be homogeneous. Specifically, they take $m-t(x)-Q(\mu)=(V_b+\varsigma\mu_b-\varsigma\mu)^+$,

where V_b, ς, and μ_b (following the notation in **?**) are all non-negative constants. (In their model the patient utility form is homogeneous and thus does not depend on the distance from the specialist. It can be shown that the service provider's optimal decisions are the same in both papers if $t(x) = 0\ \forall x$; see Section **??**.) One of our main results in this section, Theorem **??** below, still holds if we assume that the utility from treatment (or service) is given by a random variable instead of $m - t(x)$ (as long as the technical conditions (i) and (ii) in Lemma **??** in Appendix **??** hold), extending the results in **?**. By considering the quality cost, Q, as well as customer heterogeneity, we expand the growing literature on service operations. We are especially interested in the cost associated with traveling to a specialist's office, which requires considering customer heterogeneity.

3.2. Equilibrium for fixed service rate and price:

Given the specialist's choices of service rate and price, patients' decisions determine the equilibrium arrival rate. Patients have two choices, to seek treatment or not, based on their net utility $\Psi(\lambda, p, \mu, x)$. If the net utility $\Psi(\lambda, p, \mu, x)$ of a given patient from seeking treatment is non-negative, or, in other words, the benefit collected at the end of the service compensates for the expected total cost incurred in seeking the service, then the patient decides to seek treatment. Patients make the choice considering only their net utility and thus act in a self-interested manner. Patients keep joining the queue as long as their expected net utility is positive. An equilibrium is reached when no more patients have an incentive to seek treatment or can obtain a positive net utility.

Because of the special structure of the utility function Ψ, it is not difficult to see that for a given (μ, p), an equilibrium (in the sense defined in **?**; see Appendix **??** for an overview) has to have the following structure: There exists a threshold, x^*, such that patients whose distance, x, from the specialist satisfies $x \le x^*$ seek treatment and other patients do not seek treatment. Let $\lambda(\cdot,\cdot) : \mathbb{R}^2_+ \to \mathbb{R}$, where $\lambda(p, \mu)$ is the equilibrium arrival rate of such rational self-interested patients per unit time for fixed $p \ge 0$ and $\mu \ge 0$. The arrival rate under this equilibrium is given by

$$\lambda(p, \mu) = \Lambda F(x^*). \tag{2}$$

Hence, the threshold x^* can be found using the following identity:

$$x^* = \inf\{x \ge 0 : \Psi(\Lambda F(x), p, \mu, x) \le 0 \text{ and } \Lambda F(x) \le \mu\} \wedge X_m, \tag{3}$$

where the condition $\Lambda F(x) \le \mu$ ensures stability and the inf of an empty set is taken to be 0 by convention.

The intuition behind (**??**) is as follows. Assume that $\Psi(\Lambda F(x^*), p, \mu, x^*) = 0$ for some x^*. Then the patient who is at a distance x^* from the specialist is indifferent between seeking treatment

and not seeking treatment. Also, $\Psi(\Lambda F(x^*), p, \mu, x) > \Psi(\Lambda F(x^*), p, \mu, x^*) = 0$ for all $x \leq x^*$. Hence all patients located at a distance less than or equal to x^* seek treatment, so $\lambda(p, \mu)$ is the highest possible arrival rate such that all patients seeing the specialist derive a non-negative utility from seeking treatment.

3.3. Revenue-maximizing specialist:

Given patient utility function Ψ, the specialist's revenue is also directly proportional to the equilibrium arrival rate—the number of patients seen per unit time—when the specialist is paid on a fee-for-service basis. The specialist has control over the rate at which he sees patients (service rate) and the price per visit (we consider price to be exogenous in Section **??**). The specialist's choice of service rate and price automatically determines an equilibrium arrival rate based on the patient utility function (see **?**, **?**, and **?** for a similar approach in different applications). The specialist then has the revenue function $R : \mathbb{R}^2_+ \to \mathbb{R}$, defined by

$$R(p, \mu) = p\lambda(p, \mu). \tag{4}$$

The revenue-maximizing specialist will then try to choose an optimal price and service rate to obtain the maximum revenue given by

$$R^* = \sup_{p \geq 0,\ \mu > 0} R(p, \mu). \tag{5}$$

Note that the specialist is facing a complex and nonintuitive trade-off. If the specialist works faster, it will result in more appointment slots, and this in turn will lead to less congestion. This increase in the utility of an individual patient is captured through a lower expected congestion cost $c\mathbb{E}W(\mu, \lambda)$ in our model. Before walking into the clinic of the specialist, patients prefer shorter wait times. However, once meeting with the specialist face-to-face, patients prefer longer visit times with the specialist, as modeled using $Q(\mu)$. Therefore, even though faster service tends to increase the arrival rate, it will also increase the quality cost (captured by $Q(\mu)$) and hence tends to reduce the equilibrium arrival rate. The opposite effects can be seen if the specialist tries to work slowly.

Specialists usually have some degree of flexibility as to how much time they spend with each patient on average (**?**), while a minimum time with each patient is recommended and can be accounted for in Q. This is also supported by our observations that a specialist's time with the patient typically has two primary components. The first is the clinical exam and diagnostic briefing, and the second part comprises the psychological support and long-term guidance that are important when dealing with patients who suffer from incurable (chronic) disease. The latter is

in the specialist's discretion and gives support to our assumption that the specialist can alter his service rate.

Although most primary care physicians, family physicians, and pediatricians have limited bargaining power and prices are guided by medical standards determined by Current Procedural Terminology (CPT) codes and reimbursed by the insurance provider, specialists, through their associations, have sufficient say in determining the prices in the long run (**??**). In the U.S. there is a growing number of medical specialists who bypass the insurance system and charge patients directly (**??**). Outside the U.S., in countries such as Australia, France, and Finland, medical practitioners, especially specialists, are free to set their own prices for the services they offer (**???**). Therefore, modeling price as a decision variable for the specialist is a plausible approach.

3.4. Effect of the Travel Burden

We next analyze the effect of the travel burden on the equilibrium arrival rate and the optimal service rate for the revenue-maximizing specialist. We have the following proposition. (The proofs of the results in this section are placed in Appendix **??**.)

PROPOSITION 1. *Consider two travel burden functions t_1 and t_2 such that $t_2(x) = t_1(x) + a$ for some constant $a \geq 0$, for all $x \geq 0$. Let (λ_i^*, μ_i^*) denote the optimal decisions for two independent revenue-maximizing specialists treating two identical populations except for the travel burden function given by t_i, for populations $i = 1, 2$, and that the optimal revenue R_2^* (see (**??**)) for the second specialist satisfies $R_2^* > 0$. The following results hold:*[5]

(i) $\lambda_1^* \geq \lambda_2^*$, *and*

(ii) $\mu_1^* \geq \mu_2^*$.

The condition $R_2^* > 0$ is required for technical reasons. If $R_2^* = 0$, then $\lambda_2^* = 0$ is an optimal solution, so Proposition **??**(i) holds trivially.

The implication of these results is that if the distance cost or traveling cost reduces the utility of seeking treatment and hence fewer patients are interested in seeking out a specialist, then the revenue-maximizing specialist sees fewer patients. The second part of the proposition implies that quality costs decrease as the traveling cost increases. This may be interpreted in two ways. First, the specialist tries to induce demand by increasing the utility for patients by decreasing the quality cost. It can also be interpreted as the specialist trying to compensate for the travel burden by spending more time with patients. In other words, because the patients travel from far-off places or go through great difficulty in coming to the specialist, he tries to spend more time with them to make them feel better, thereby increasing their net utility.

[5] The results also hold if $t_1(x) = 0$ and $t_2(x) \geq 0$ is a travel burden function that is a constant or satisfies Assumption **??**. In addition, the case $t_1(x) = 0$ corresponds precisely to the analysis in **?**.

3.5. Welfare-maximizing specialist

Next, we compare the optimal decisions of a revenue-maximizing specialist with those of a welfare-maximizing specialist. First, we present the details of our model for the welfare-maximizing specialist, following **?**.

If the welfare-maximizing specialist cannot serve all patients, he would prefer to choose the patients with the greatest benefits from treatment. With other parameters remaining the same, the closer the patients are located, the greater the benefits they receive. Hence the welfare-maximizing specialist needs to determine the maximum distance (and he would serve all the patients located closer than this maximum level) and choose the service rate. Therefore, we can transform the problem to choosing an arrival rate (instead of choosing the maximum distance), as we explain next.

Let λ_x denote the total arrival rate of patients whose distance is at most x. Then

$$\lambda_x = \Lambda F(x). \tag{6}$$

By inverting we can find x_λ, the maximum distance of an arriving patient, given the total arrival rate λ and assuming that only patients at a distance less than x_λ will arrive. Specifically, $x_\lambda : [0, \Lambda] \to [0, X_m]$ is given by

$$x_\lambda = \begin{cases} F^{-1}(\lambda/\Lambda), & \text{for } \lambda \in (0, \Lambda) \\ X_m, & \text{if } \lambda = \Lambda \\ 0, & \text{if } \lambda = 0 \end{cases}. \tag{7}$$

Then $V : [0, \Lambda] \to \mathbb{R}$, the cumulative benefit for all patients obtaining service net of travel burden, as a function of the total arrival rate λ, is given by

$$V(\lambda) = \Lambda \int_0^{x_\lambda} (m - t(x))\, f(x)\, \mathrm{d}x. \tag{8}$$

The total utility $U : [0, \Lambda] \times \mathbb{R} \to \mathbb{R}$ per unit time obtained by all the patients, if λ patients seek service per unit time and the specialist employs a service rate μ per unit time, is given by

$$U(\lambda, \mu) = \begin{cases} V(\lambda) - \lambda \left(Q(\mu) + \frac{c}{\mu - \lambda} \right), & \text{if } \mu > \lambda, \text{ and } \lambda \in [0, \Lambda] \\ 0, & \text{otherwise} \end{cases}, \tag{9}$$

where the condition $\mu > \lambda$ again ensures stability. The objective from the welfare-maximizing angle would be to maximize the total utility of all patients per unit time. Thus the objective would be to find

$$U^* = \sup_{\lambda \geq 0,\ \mu > 0} U(\lambda, \mu). \tag{10}$$

In summary, a welfare maximizer needs to strike a balance between serving more patients and managing wait times in a cost-effective way. With each additional patient, the specialist's workload goes up, so if the average service rate remains the same, the congestion cost increases for all patients. Each additional patient therefore increases the cost for existing patients and hence the overall system. A welfare maximizer will consider the negative externality caused by the arrival of each additional patient. Prices charged will simply be transfers, as the specialist is also part of this joint system that has patient welfare in mind. For welfare maximization we allow non-positive prices, as patients can be compensated in different ways for receiving treatments. Hence the objective from the welfare-maximizing angle will be to maximize the total utility of all patients, ignoring the price transfers, as in (**??**).

Comparing decisions for the two different objectives: We next compare the optimal decisions of the revenue-maximizing specialist and the welfare-maximizing specialist.

THEOREM 1 **(Comparison of the two specialists)**. *Let (λ^*, μ^*) denote the optimal decisions for the revenue-maximizing specialist, and let (λ_s^*, μ_s^*) denote the optimal decisions for the welfare-maximizing specialist.*

i. The optimal equilibrium arrival rate for the welfare-maximizing specialist is never lower than that for the revenue-maximizing specialist; that is, $\lambda_s^* \geq \lambda^*$, if $R^* > 0$.

ii. The optimal service rate for the welfare-maximizing specialist is never lower than that for the revenue-maximizing specialist; that is, $\mu_s^* \geq \mu^*$, if $\lambda_s^* > 0$ and $\lambda^* > 0$.

iii. If $t(x)$ is constant and $\lambda^* > 0$, then $\lambda_s^* = \lambda^*$ and $\mu_s^* = \mu^*$.

From Theorem **??**, a revenue-maximizing specialist sees fewer patients per unit time than a welfare-maximizing specialist does. Just like a monopolist, the revenue-maximizing specialist earns higher revenue by charging a higher price for the fewer patients he sees. However, the higher revenue does not compensate for the lost utility, as some patients opt out of treatment because of the combined effect of the higher prices and the travel burden. Hence, the total welfare is lower. We note here that this part of the proposition is similar to what **?** found, and our proof follows similar lines as well, but because of the addition of the service quality effect ($Q(\mu)$) the extension is not straightforward. Moreover, it is obvious that the first part of the result is also true if $\lambda^* = 0$.

From the second part of the proposition we observe that the revenue-maximizing specialist sees patients at a slower rate than the welfare-maximizing specialist. Even though the former spends more time with patients, it might not be socially optimal to work at such a slow rate. Also, even though the service rate increases, the specialist sees more patients in the welfare-maximizing case.

In fact, we show in Appendix **??** that the *utilization*, hence the congestion cost $(c/(\mu-\lambda))$, is not lower than it is in the case of a revenue-maximizing specialist; see Remark **??** for details.

The third part of the proposition states that if customers are homogeneous in their travel burden, as in **?**—i.e., $t(x)$ is a constant, then the optimal actions of the two specialists will be identical. This result is similar to that in **?**, but they studied the case when the service provider cannot alter the service rate. We also note here that when customers are homogeneous the specialist sets a price such that all patients receive zero utility. Otherwise, more patients will be expected to join, as the utility is positive.

3.6. Exogenous price setting

In this section, we analyze the case in which visit prices are determined exogenously by an external regulator or by a health insurance company and may be outside the control of the agency (which might have preferred to act differently), for example, due to political reasons. We use the model we introduced in Section **??** for the patients and providers, but now we assume that the price is fixed at p. Therefore, the specialist has control over only the average visit length when he sees patients at the clinic or, in other words, the service rate. The choice of the service rate by the specialist automatically determines the equilibrium arrival rate in a way similar to the one described in Section **??**.

Although the case with fixed prices may seem like a special case of the more general case we considered in Section **??**, surprisingly Theorem **??** does not hold anymore. We illustrate this using a case in which the cost and utility functions have a special structure. We make the following simplifying assumptions: $t(x)=t_o x$, where t_o is the transportation cost per unit distance and x is the distance from the specialist to the patient; $Q(\mu)=\delta\mu$, where μ is the service rate and $\delta>0$ represents the proportionality constant; $f(x)=1/M$ for $x\in[0,M]$. For notational simplicity, we also take $\Lambda=M/t_0$. The price that maximizes the specialist's revenue in (**??**) is denoted by p^*, and we let p_s^* denote the exogenous price that is derived from the optimal solution of (**??**)—that is, p_s^* satisfies $\lambda(p_s^*,\mu_s^*)=\lambda_s^*$. We use $\lambda^*(p)$ and $\mu^*(p)$ to denote the arrival and service rates for a revenue-maximizing provider when the exogenous price is set to p. We next prove that the conclusions in Theorem **??** are not valid if the fixed price is not high enough.

PROPOSITION 2. *If the exogenous price $p<p_s^*$, then*

a. *the optimal equilibrium arrival rate for the revenue-maximizing specialist would be higher than that for the welfare-maximizing specialist—that is, $\lambda^*(p)>\lambda_s^*$;*
b. *the optimal service rate for the revenue-maximizing specialist would be higher than that for the welfare-maximizing specialist—that is, $\mu^*(p)>\mu_s^*$.*

If $p > p_s^*$, then $\lambda^*(p) < \lambda_s^*$ and $\mu^*(p) < \mu_s^*$, similar to Theorem **??**.

We observe that there exists a threshold for the fixed price per visit, p, offered to the specialist by a third party. Above this threshold, p_s^*, the optimal arrival and service rates for a revenue-maximizing specialist would be lower than those for a welfare-maximizing specialist, which is similar to the result in Theorem **??**. From Proposition **??**, this threshold is below that of the price that a revenue-maximizing specialist would set (the endogenous case). Since one would expect the exogenous price, p, to be less than p^*, Proposition **??** means that there are cases when even if the price is determined by a third party ($p_s^* < p < p^*$), the revenue-maximizing specialist would work more slowly than the welfare-maximizing specialist. We further illustrate Proposition **??** using a numerical analysis in Appendix **??**.

4. Analyzing the operational impact of telemedicine

We now model the implications of implementing telemedicine for both utility-maximizing patients and revenue-maximizing specialists. We assume that the specialist may choose to offer two modes of treatment, telemedicine and in-person. A patient then has three choices: (1) in-person mode, (2) telemedicine mode, and (3) no treatment. A patient's choice will depend on the utility of each option.

For most medical conditions, the physician's vision is the primary medium for diagnosis. An interactive video-conferencing system thus can enable a specialist to carry out a significant portion of the assessments required in a typical one-on-one visit. However, patients might still need to travel to the specialist's location for some clinical assessments, laboratory procedures, and emergency situations. Thus, not all visits can be done via telemedicine. With a slight abuse of terminology, we refer to a consultation as an in-person visit if the patient travels to see the specialist and as a telemedicine "visit" if the consultation is done remotely using telemedicine technology.

In Section **??**, we introduce a model in which the specialist dedicates the same amount of time (on average) to each in-person and telemedicine visit and charges the same price for both. Then, in Section **??**, we consider the case in which the specialist can choose different service rates and prices. Although we were able to analyze the former model in detail, the equilibrium behavior of patients is much more complex for the latter model, so we make additional assumptions and offer numerical results after we present a method to identify the equilibrium.

4.1. Optimal decisions for a specialist offering the telemedicine mode

We start our analysis of the effect of telemedicine on the specialist's and patients' decisions by considering a special case in which the specialist does not differentiate between telemedicine and

in-person visits but chooses the same service rate μ and price p for all patients. From the specialist's perspective, the patient's mode of treatment does not really matter (since the service rate and price are identical for both modes). However, from a patient perspective, the patient's utility is dependent upon the mode of treatment due to different travel burdens.

We use a model similar to that in Section **??** to capture patient utility. We assume that a patient located at a distance x from the specialist gains utility $m_i - t(x)$ from an in-person visit and utility $m_t - t(0)$ from a telemedicine visit, for two positive constants m_i and m_t. (Recall that $t(x)$ denotes the travel burden for a patient located at a distance x from the specialist and that it is assumed to be strictly increasing.) Term $t(0)$, being a constant, can be adjusted in the factors m_i and m_t, so we take $t(0) = 0$ without loss of generality. (We assume that the conditions in Appendix **??** continue to hold.) Similar to (**??**), for a given total arrival rate λ, price p, service rate μ, and distance of the patient x, we define the patient's utility $\Psi_i(: [0, \Lambda] \times \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R})$ for the in-person mode by

$$\Psi_i(\lambda, p, \mu, x) = m_i - t(x) - Q(\mu) - \beta p - \frac{c}{\mu - \lambda}. \tag{11}$$

The main difference between Ψ_i and Ψ in (**??**) is that λ is the total arrival rate, including those who chose the telemedicine mode. As noted earlier, not all the care provided in the office can be delivered by telemedicine sessions. Accordingly, for telemedicine patients we define the clinical feasibility, $\alpha \in (0, 1)$ and $(1 - \alpha)$, representing the fraction of visits via telemedicine and in person respectively. If the patient located at distance x from the specialist chooses the telemedicine mode and if s is the setup cost involved in telemedicine (further explained below), the patient utility $\Psi_t(: [0, \Lambda] \times \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R})$ is given by

$$\Psi_t(\lambda, p, \mu, x) = \alpha m_t + (1 - \alpha)(m_i - t(x)) - s - Q(\mu) - \beta p - \frac{c}{\mu - \lambda}. \tag{12}$$

We use $s \in \mathbb{R}_+$ to denote the amortized cost of setting up telemedicine visits, such as setting up an Internet connection, installing a webcam (or other video-conferencing facility), and fulfilling other hardware and software requirements; see **?**. It also includes the operating cost and any cost related to on-site technical support that may be needed for the initial few visits. The other terms are similar to those in Section **??**.

Objective function and decisions: Let $\lambda_t(p, \mu)$ denote the equilibrium arrival rate of telemedicine visits and $\lambda_i(p, \mu)$ the equilibrium arrival rate of in-person visits per unit time, whose existence we prove below, for a given price p and service rate μ. Most of the cost incurred by a specialist will be fixed in nature (the cost of the facility, staff, and so on). Most specialists devote a certain portion of their time to seeing patients. Hence the cost is likely to depend on the aggregate

time and not the time spent with an individual patient or type of visit. In other words, we assume that the cost of implementing telemedicine is negligible. If the specialist chooses a price p and a service rate μ, then the revenue function $\bar{R}:\mathbb{R}_+^2 \to \mathbb{R}$ for the specialist is defined by

$$\bar{R}(p,\mu) = \begin{cases} p(\lambda_i(p,\mu)+\lambda_t(p,\mu)), & \text{if } p \geq 0, \\ 0, & \text{otherwise} \end{cases}. \tag{13}$$

The revenue-maximizing specialist will try to choose an optimal price and service rate to suit his objective. The objective function for the specialist is then given by

$$\bar{R}^* = \sup_{p\geq 0,\ \mu>0} \bar{R}(p,\mu). \tag{14}$$

Equilibrium for a specialist offering both modes: We next establish the existence of the equilibrium when patients have a choice between in-person and telemedicine modes, based on the definition of a Nash equilibrium for a continuum of players as in **?**.

PROPOSITION 3. *Given the payoff functions* (**??**), (**??**), *and* (**??**) *for in-person and telemedicine patients and for the revenue-maximizing specialist, respectively, there exists a unique equilibrium.*

In proving Proposition **??**, we also determine the potential equilibrium outcomes in closed form, which are instrumental in our analysis of telemedicine's impact on welfare in the next section. Also, we observe that in equilibrium some patients may or may not seek telemedicine mode depending on the utility they receive from different modes of treatment.

Unfortunately, the proof of Proposition **??** cannot be extended to the case when the specialist's choices are different for the two modes of treatment. This is because the provider's objective function in (**??**) is separable in a certain sense only when service rates and price are the same for both modes.

REMARK 2. Throughout this section we assume price to be a decision variable for the specialist for two reasons. First, most telemedicine services are currently not reimbursable, and the specialist sets the price in these cases (**?**, **?**), and we are interested in the long-run equilibrium actions of providers. Second, when prices are exogenous, the equilibrium outcome depends on the price; recall Section **??**. However, we verify numerically that the insights generated when prices are exogenous and fixed are, in principle, similar if prices are low enough, since the lower exogenous prices will tend to drive the optimal service rates higher; see Appendix **??**.

Economic feasibility of telemedicine: In this section we explore the effect of offering the telemedicine mode on the specialist's optimal actions. Let $\tilde{\lambda}_i$ and $\tilde{\lambda}_t$ denote the optimal equilibrium arrival rates of in-person and telemedicine patients and $\tilde{p}$ and $\tilde{\mu}$ denote the optimal equilibrium price and service rate. Recall that we use λ^* and μ^* to denote the equilibrium arrival rate when a revenue-maximizing provider does not offer the telemedicine mode.

THEOREM 2. *Assume that* $\lambda^* > 0$. *If*

$$s < \alpha \left(m_t - m_i + t \left(F^{-1} \left(\lambda^* / \Lambda \right) \right) \right), \tag{15}$$

then

i) the optimal equilibrium total arrival rate for a specialist offering both modes of service is greater than that for a specialist offering only the in-person mode —that is, $\tilde{\lambda}_t + \tilde{\lambda}_i > \lambda^*$*; and*

ii) the optimal service rate for a specialist offering both modes of service is greater than that for a specialist offering only the in-person mode —that is, $\tilde{\mu} > \mu^*$.

The result gives a simple necessary condition to check the economic feasibility of telemedicine based on the patient (located at λ^*/Λ) who is indifferent between the in-person mode of treatment and telemedicine mode of treatment. The condition in (**??**) defines a threshold for the setup costs of telemedicine based on the direct relative treatment benefit telemedicine visits provide ($m_t - m_i$) and the reduction in travel burden (proportional to α and t). The significance of the result is that the condition can be checked based on the parameters that are observed before the introduction of telemedicine, except m_t, the treatment benefit from telemedicine.

By part (ii) of Theorem **??**, the specialist sees more patients at a faster rate. Thus, the results from our model support the hypothesis that telemedicine increases patient access to specialists; see **?**. Interestingly, the increased access comes from increasing the "efficiency" in the system by means of an increased service rate. Recent work by **?** and **?** also offer empirical support that telemedicine could increase specialist efficiency through shortened visit times. However, it is to be noted that those patients who continue to choose the in-person mode of treatment will experience a reduced net utility because of the faster service rate.

Effect of telemedicine on total welfare: Although the introduction of telemedicine increases the arrival and service rates, its exact effect on total welfare is unclear. Also, we cannot use Theorems **??** and **??** to reach a conclusion because of the differences in the current model when compared to that studied in Section **??**. Therefore, we next analyze how total welfare is impacted by telemedicine and show that it increases under certain conditions.

First, we define the total welfare when some patients choose telemedicine. Let x_{ID} denote the location of the patient who is indifferent between the two modes of treatment; that is,

$$x_{ID} = \inf\{x \geq 0 : \alpha(m_i - t(x) - m_t) + s \leq 0\}. \tag{16}$$

Due to the independence of x_{ID} from μ and p, if the specialist chooses arrival rate λ, then those patients located at a distance between 0 and $(x_{ID} \wedge F^{-1}(\lambda/\Lambda))$ receive higher utility from the in-person mode, and those located between $(x_{ID} \wedge F^{-1}(\lambda/\Lambda))$ and $F^{-1}(\lambda/\Lambda)$ receive higher utility from the telemedicine mode (see Lemma **??** for more details). Therefore, the total welfare, if the specialist chooses to serve an arrival rate of λ and uses a service rate of μ, is given by

$$U_d(\lambda, \mu) = \begin{cases} V_d(\lambda) - \lambda\left(Q(\mu) + \frac{c}{\mu - \lambda}\right), & \text{if } \mu > \lambda, \text{ and } \lambda \in [0, \Lambda] \\ 0, & \text{otherwise} \end{cases}, \tag{17}$$

where

$$V_d(\lambda) = \Lambda\left[\int_0^{x_{ID} \wedge F^{-1}(\lambda/\Lambda)} (m_i - t(x))\, f(x)dx + \int_{x_{ID} \wedge F^{-1}(\lambda/\Lambda)}^{F^{-1}(\lambda/\Lambda)} (\tilde{m} - (1-\alpha)t(x))\, f(x)dx\right], \tag{18}$$

and $\tilde{m} = \alpha m_t + (1-\alpha)m_i$. Let $\tilde{U}_d (:= U_d(\tilde{\lambda}_i + \tilde{\lambda}_t, \tilde{\mu}))$ denote the resulting total welfare in equilibrium for a revenue-maximizing provider who offers the telemedicine mode, and recall that $U(\lambda^*, \mu^*)$ is the total welfare in equilibrium for a revenue-maximizing provider's optimal actions, λ^* and μ^*, *without telemedicine*.

THEOREM 3. *If* (**??**) *holds, then the total welfare increases with the introduction of telemedicine; that is,* $\tilde{U}_d \geq U(\lambda^*, \mu^*)$.

4.2. Treatment model using distinct service rates and prices

Next we consider a more general model for the specialist's approach to the telemedicine mode based on our observations in practice, where the specialist allocates a proportion, r, of his capacity to in-person visits and the rest, $1 - r$, to telemedicine visits. This is typical in certain specialties where the specialist only sees telemedicine patients during certain hours each day or on certain days of the week. The specialist in effect provides the two modes of service almost independently. However, recall that patients using the telemedicine mode still need in-person visits for a fraction of their visits, tying these two modes together. In addition, we assume that the provider may choose different prices and service rates for the two modes.

In this case the model becomes analytically intractable, and we were unable to obtain the potential equilibrium outcomes in a closed form, so we cannot generalize Theorems **??** and **??**. Hence we carry out a numerical analysis under the following streamlining assumptions.

Assumptions:

C.1 The travel burden is linear in distance, $t(x) = t_o x$, where t_o is the transportation cost per unit distance and x is the distance from the specialist to the patient.

C.2 The service quality function is linear, $Q(\mu) = \delta\mu$, where μ is the service rate and $\delta > 0$ represents the proportionality constant.

C.3 Patients are uniformly distributed with the specialist located at $x = 0$; that is, $f(x) = 1/M$ for $x \in [0, M]$.

We also take the total arrival rate if the specialist decides to serve all the patients $\Lambda = M/t_0$ for analytical simplicity. The specialist has five variables to optimize: the service rates and the prices for each mode, and the proportion of time dedicated to each mode. Because total enumeration is not practical due to the size of the problem, we developed a method explained in **?** to identify the optimal actions of a specialist given r. We then used a numerical search algorithm to find the optimal r.

As a base case, we considered the following values for the various parameters in our model: the proportionality constant for the service quality function, $\delta = 1$; the patient reward from an in-person visit, $m_i = 60$; the patient reward from a telemedicine visit, $m_t = 60$; the transportation cost per unit distance, $t_o = 10$; the co-insurance rate, $\beta = 0.1$; the cost to the patient for setting up telemedicine visits, $s = 10$; the clinical feasibility of telemedicine, $\alpha = 0.75$; and the opportunity cost per unit time for patients, $c = 5$. We also carried out a sensitivity analysis by allowing m_t and t_o to vary from 0 to 90 and 0 to 30 respectively with 30 equal increments, keeping other values as in the base case. Similarly, we considered 30 different values for α from 0.01 to 1 with equal increments and for s from 0 to 45 with equal increments. We mainly focus on the sensitivity of the results with the parameters α and t_0 as these two parameters define the nature of telemedicine visits. The results of additional numerical results are presented in Appendix **??** for brevity, and we refer to them while we explain the reasons behind our general observations.

Our objectives in this section are as follows. We first would like to check whether our results in Theorems **??** and **??** will still hold in this new setting—i.e., the effect of telemedicine on total coverage (see Section **??** below) and social welfare (Sections **??** and Section **??** below). In addition, the numerical analysis will help us in understanding the extent of the impact of telemedicine on the specialist's revenue. With separate prices for in-person and telemedicine visits in this setting, we will also compare the total patient surplus (that is, the sum of the utility of all patients) before and after telemedicine, something that we could not do in Section **??**.

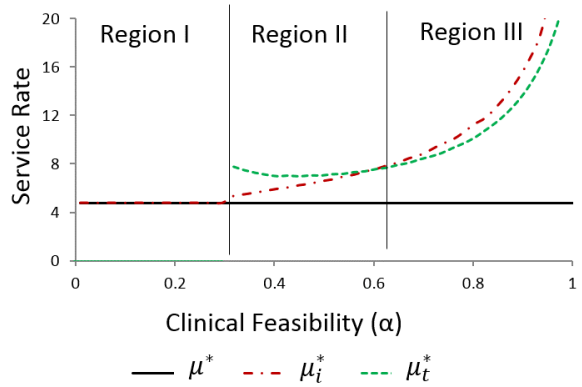
4.2.1. Optimal actions and resulting revenue: We first focus on the optimal actions and the optimal revenue after the introduction of telemedicine. Figure **??** compares the optimal service rates for the specialist and the equilibrium arrival rates at optimal values before telemedicine (μ^* and λ^*) and after telemedicine (μ_i^* and λ_i^* for in-person visits and μ_t^* and λ_t^* for telemedicine visits) as functions of α and t_0, respectively. The vertical lines in the following figures are used to separate the different regions. Region I is when the specialist does not introduce telemedicine in equilibrium. The other regions are explained below as and when needed.

The optimal service rates and arrival rates are higher than they were before telemedicine for high enough α and t_0 ($\alpha > 0.3$ in Figures **????** and **????** and $t_0 > 1$ in Figures **????** and **????**). Thus our results in Theorem **??** hold in this case as well. The service and arrival rates also increase with the clinical feasibility of telemedicine (α) (Figures **????** and **????**), and they decrease with the travel burden (t_0) (Figures **????** and **????**).

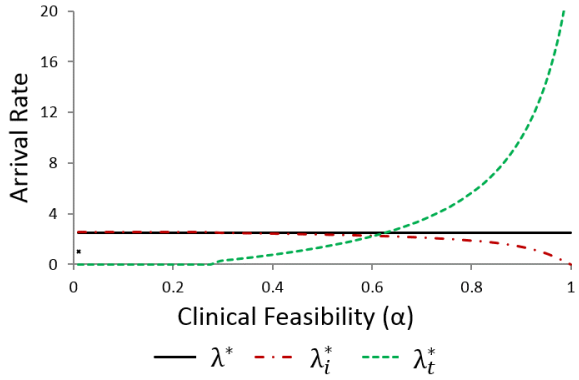
The optimal service rate for telemedicine visits can be either higher (Region II, Figure **????**) or lower (Region III, Figure **????**) than the optimal service rate for in-person visits. Basically, the specialist plays with the two levers, price and service rate, to maximize his revenue. For low values of α, the price for telemedicine visits is also relatively lower (see Appendix **??**). Hence the service rates are higher to optimize revenue. For higher values of α, prices are higher, and this increase is compensated by a decrease in the service rate.

Figures **????** and **????** compare the optimal revenue for the specialist before telemedicine (R^*) and after telemedicine ($R^*\ TM$) (both on the left axis), and also show the proportion of time spent on in-person visits (r) (on the right axis), as functions of the clinical feasibility of telemedicine, α (Figure **????**), and the transportation cost, t_0 (Figure **????**). As the clinical feasibility of telemedicine (α) increases, the proportion of time spent on in-person visits (r) decreases. In other words, as telemedicine becomes more clinically feasible and a higher proportion of visits can be done virtually, the specialist will also find it beneficial to adopt telemedicine. From a policy perspective, telemedicine may not make sense for chronic conditions where α is relatively small. In Figure **????** the specialist will only offer the telemedicine mode beyond a certain threshold ($\alpha = 0.3$). In addition, both R^*, the specialist's revenue before telemedicine, and $R^*\ TM$, the specialist's revenue after telemedicine, decrease with the transportation cost (t_0).

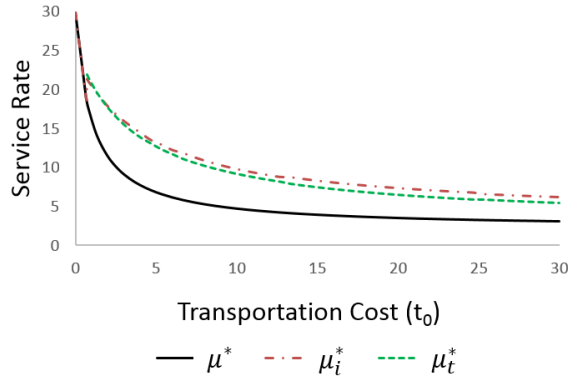
In summary, specialists' productivity and revenue increase, but so will their utilization (see Appendix **??**) with the introduction of telemedicine. Patients can expect to enjoy a reduction in their travel burden, which increases their utility, resulting in greater patient access (higher arrival rates). On the other hand, patients can also expect increased congestion (waiting times for



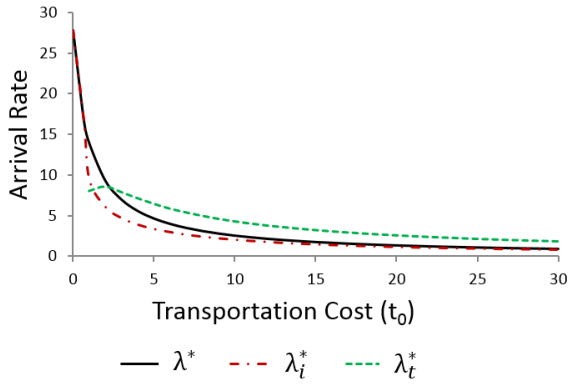
(a) Optimal Service Rates vs. Clinical Feasibility of Telemedicine ($\boldsymbol{\alpha}$): $m_t = 60; t_0 = 10$



(b) Arrival Rates vs. Clinical Feasibility of Telemedicine ($\boldsymbol{\alpha}$): $m_t = 60; t_0 = 10$



(c) Optimal Service Rates vs. Transportation Cost ($\boldsymbol{t_0}$): $\alpha = 0.75; m_t = 60$



(d) Arrival Rates vs. Transportation Cost ($\boldsymbol{t_0}$): $m_t = 60; t_0 = 4$

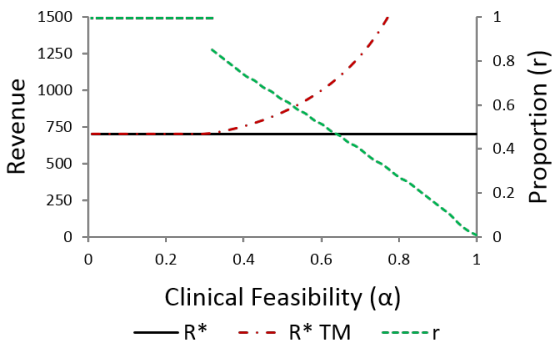
Figure 1 **Analyzing the optimal service rates (μ^*, μ_i^*, and μ_t^*) and the optimal arrival rates (λ^*, λ_i^*, and λ_t^*):** $\delta = 1; m_i = 60; \beta = 0.1; s = 10; c = 5$

appointments; see Appendix **??**), shorter visits online or in person, and busier specialists, which tend to decrease their utility. We also find that if the perceived quality of telemedicine visits is comparable to that of in-person visits, the average price per visit decreases with the introduction of telemedicine (see Appendix **??**).

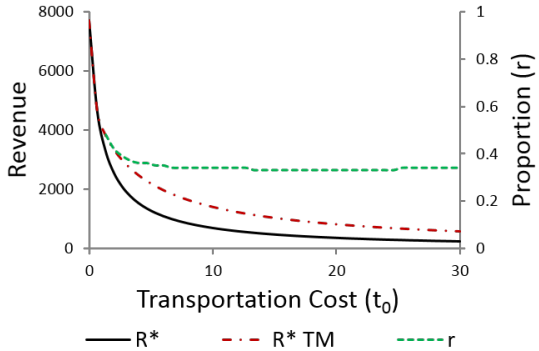
4.2.2. Effect on total welfare: As we demonstrated in the previous section, even though the travel burden decreases for patients choosing the telemedicine mode, the higher service quality cost and the higher congestion cost reduce the utility for some patients. It is therefore not clear what happens to the total welfare[6] and whether Theorem **??** still holds in this case.

Figures **????** and **????** compare the total welfare for a revenue-maximizing specialist before

[6] The total welfare is defined in a manner similar to (**??**) by ignoring the prices paid by patients. The main difference is that we account for the utilities of both in-person and telemedicine patients.



(a) Optimal Revenues (R^* and R^*TM) and Proportion of Time Spent on In-person Visits ($\boldsymbol{r}$) vs. Clinical Feasibility of Telemedicine ($\boldsymbol{\alpha}$): $m_t = 60; t_0 = 10$



(b) Optimal Revenues (R^* and R^*TM) and Proportion of Time Spent on In-person Visits ($\boldsymbol{r}$) vs. Transportation Cost ($\boldsymbol{t_0}$): $\alpha = 0.75; m_t = 60$

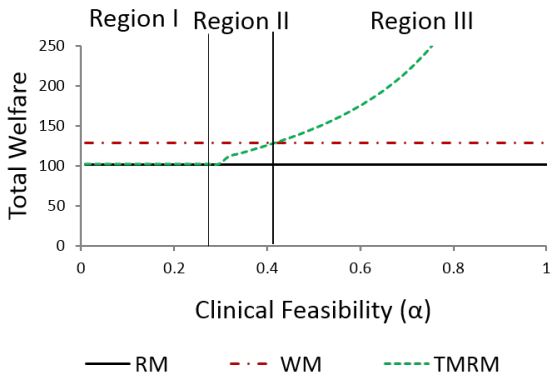
Figure 2 **Analyzing the optimal revenues (R^* and R^*TM) and proportion of time spent on in-person visits (r):** $\delta = 1; m_i = 60; \beta = 0.1; s = 10; c = 5$

telemedicine (RM), the total welfare for a welfare-maximizing specialist before telemedicine (WM), and the total welfare for a revenue-maximizing specialist after telemedicine (TMRM) as functions of the clinical feasibility of telemedicine, α (Figure **????**), and the travel burden, t_0 (Figure **????**). There is no difference between RM and WM when customers are homogeneous ($t_0 = 0$) (not shown in Figure **????** to emphasize the difference between the welfares when $t > 5$), but the difference increases with the degree of heterogeneity, as seen in Figure **????**.

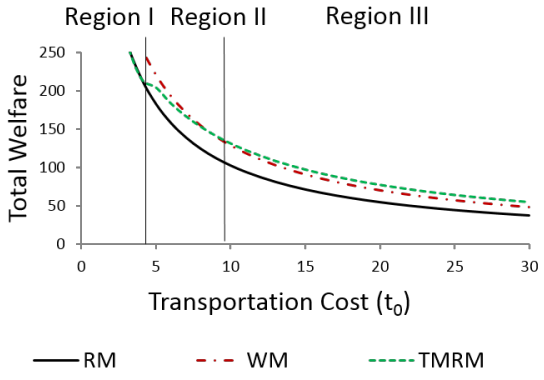
Surprisingly, the total patient welfare increases when the specialist offers the telemedicine mode (in Regions II and III in Figures **????** and **????**), even for a revenue-maximizing specialist. Thus, even though congestion costs increase for patients who were undergoing treatment before telemedicine was made available, the reduced travel burden for the existing patients and the added welfare through patients who were not treated before more than compensates for this relative welfare loss. Theorem **??** therefore holds in this case.

The total welfare under a revenue-maximizing specialist who introduces telemedicine (TMRM) is more than the welfare under a welfare-maximizing specialist who does not offer telemedicine (WM) if the transportation cost is sufficiently high (Region III, Figure **????**) or if the clinical feasibility of telemedicine is not too low (Region III, Figure **????**). When the clinical feasibility of telemedicine or the travel burden is in Region II, the total welfare under telemedicine is more than that of the revenue-maximizing specialist but less than that of the welfare-maximizing specialist.

4.2.3. Effect on total patient surplus: In this section, we analyze the impact of telemedicine on total patient surplus. The total patient surplus is defined similarly to the total



(a) Total Welfare vs. Clinical Feasibility of Telemedicine ($\boldsymbol{\alpha}$): $\delta = 1; m_i = 60; t_0 = 10; \beta = 0.1; s = 10; m_t = 60; c = 5$



(b) Total Welfare vs. Transportation Cost ($\boldsymbol{t}_0$): $\delta = 1; m_i = 60; m_t = 39; \beta = 0.1; s = 10; \alpha = 0.75; c = 5$

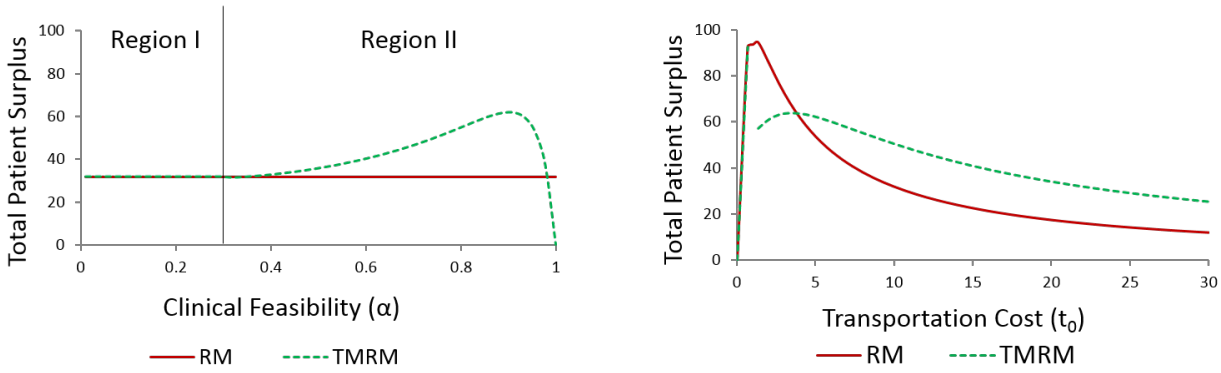
Figure 3 Analysis of total patient welfare under a welfare-maximizing specialist before the introduction of telemedicine (WM) and a revenue-maximizing specialist before (RM) and after the introduction of telemedicine (TMRM)

welfare in (**??**), but it includes the prices paid by patients to the specialist.

The introduction of telemedicine has different impacts on different patients depending on their distance from the specialist. Even though prices fall after the introduction of telemedicine (see Appendix **??**), the higher service quality cost and the higher congestion cost reduce the utility for some patients. Even with an increase in total welfare, the reduced prices may still be too high for some patients to see an increase in the total surplus. Telemedicine is thus beneficial only for patients located at some distance from the specialist for whom the reduction in the travel burden and price more than compensates for the higher congestion cost. (See Figure **??** in Appendix **??** for how a patient's utility varies with distance.)

We next turn to total patient surplus. Figures **????** and **????** compare the total patient surplus for a revenue-maximizing specialist before telemedicine (RM) and after telemedicine (TMRM) as functions of the clinical feasibility of telemedicine, α (Figure **????**), and the transportation cost per unit distance, t_0 (Figure **????**). Region I is when the specialist does not offer the telemedicine mode; $\alpha \leq 0.3$ in Figure **????** and $t_0 < 1$ in Figure **????**.

As telemedicine becomes more and more feasible (higher α), the total patient surplus increases and is higher than the total utility before the introduction of telemedicine; see Figure **????**. When α is very high ($\alpha \geq 0.9$ in Figure **????**), however, the total patient surplus starts decreasing, and it can become lower than it was before telemedicine ($\alpha \geq 0.98$ in Figure **????**). This is because as α increases more patients shift to the telemedicine mode and the heterogeneity among patients decreases as well (a higher α reduces the need for travel and hence implies a reduced travel burden



(a) Total Utility vs. Clinical Feasibility of Telemedicine ($\boldsymbol{\alpha}$): $\delta = 1; m_i = 60; m_t = 60; \beta = 0.1; t_0 = 10; s = 10; c = 5$

(b) Total Utility vs. Transportation Cost ($\boldsymbol{t_0}$): $\delta = 1; m_i = 60; m_t = 60; \beta = 0.1; s = 10; \alpha = 0.75; c = 5$

Figure 4 Analysis of total patient surplus under a revenue-maximizing specialist before the introduction of telemedicine (RM) and after the introduction of telemedicine (TMRM)

for these patients). A high α scenario therefore gets closer to the situation when $t_0 = 0$, where the specialist absorbs all the patient surplus (see Theorem **??**), by setting an appropriate price and service rate, and therefore results in a total utility close to zero.

The total patient surplus increases with t_0 at first ($t_0 < 1$) and then decreases; see Figure **????**. For low values of t_0 ($t_0 \leq 4$), the total patient surplus is lower with telemedicine due to the increase in service rates (see Figure **????**) and hence higher quality costs. But above this threshold the total patient surplus is higher with telemedicine, as the lower travel burden and increased access compensate for the higher quality costs. Thus, as the travel burden increases, not only is the total welfare greater with telemedicine (Figure **????**), the total patient surplus is higher with telemedicine as well.

5. Conclusions and key policy implications

With the rapid deployment of digital technologies, more and more specialists are looking at the potential of telemedicine in treating some or all of their patients. In this work, we analyze the operational impact for a specialist implementing telemedicine technology for the care of patients suffering from chronic conditions who are arriving periodically at the clinic from different locations. Our research extends analytical results on service interactions to the general case of heterogeneous customers for both revenue-maximizing and social welfare-maximizing service providers. We characterize the impact of patient heterogeneity on the specialist's price and service rate decisions, where the heterogeneity could be due to different morbidity or the travel burden each patient faces. We also consider the case when specialist fees (price) are fixed exogenously and its impact on the specialists' strategic behavior.

We then apply these new results in examining the economic and operational implications of implementing telemedicine for the care of chronic patients. The model especially allows us to explain how already busy specialists would still be able to accommodate additional telemedicine visits and to analyze the impact of introducing telemedicine on overall coverage, service quality, specialist productivity and income, cost of care, and patient welfare.

We obtain the following important managerial insights for the benchmark case—i.e., the operating equilibrium *before* telemedicine is introduced. We show when it would become optimal, even for physicians who are being paid on a fee-for-service basis, to spend more time on average with each patient as the travel burden increases. By doing so these physicians could end up with fewer patients per day and fewer billable visits, yet it would be optimal for them to provide their patients a better clinical experience, thereby partially recognizing their travel burden. We prove that this trend becomes more and more pronounced with an increase in the patient's travel burden. Second, under fairly general conditions of patient heterogeneity, we prove that the arrival rate of patients (or the workload) at a revenue-maximizing specialist is always lower than (or equal to) that at a welfare-maximizing specialist. On the other hand, the former spends more time with each patient than the latter. Our analysis explains why the congestion costs for the revenue-maximizing specialist would always be lower than they are for the welfare-maximizing specialist. Thus the revenue-maximizing specialist will work more slowly and treat fewer patients than the welfare-maximizing specialist, making those providers more easily accessible.

While high-end specialists can freely set their fees, others may have to accept the fees set by third-party payers. In this case we identify the threshold price above which the revenue-maximizing specialist would work more slowly than the welfare-maximizing specialist. Our results also identify the feasible range of fees determined by third-party payers and the exogenous price that results in welfare maximization.

Using non-atomic games to model patient choices for care modalities, we prove that with the introduction of telemedicine technology, a patient's strategic choice falls into one of four mutually exclusive outcomes in equilibrium: Existing patients who were treated in person choose to continue with in-person visits or switch to virtual visits, new patients (for instance, those who live farther away) now join the clinic to be seen via telemedicine, and patients (for instance, those who live even farther away) who were not treated before choose to stay untreated by that specialist. With the introduction of telemedicine, the clinic will expect to see an increase in the arrival rate of patients. This will result in higher congestion costs for the patients, increased waiting times for

appointments, shorter (online or face-to-face) visits, and busier providers. Despite all these costs, we prove that social welfare increases even for the patients served by revenue-maximizing specialists.

Our results also indicate that telemedicine benefits the specialist physicians, as they enjoy both higher productivity and higher revenue. When in-office visits provide similar or superior value as compared to telemedicine visits, in an unregulated system the optimal uniform fees charged by the specialists (for both in-office and telemedicine visits) would be lower than the fees levied before telemedicine. Our numerical experiments illustrate how the introduction of telemedicine will benefit patients, providers, and third-party payers. We also show that as the travel burden increases, specialists decrease their service rate and yet their utilization decreases as well.

While the clinical efficacy of using telemedicine for a host of conditions has been proven in prior clinical studies by us and others, there still remains a host of public policy and economic issues that prevent it from being a widely accepted clinical practice. Our research highlights some very interesting policy implications regarding some of the administrative barriers to the implementation of telemedicine for treating chronically ill patients (**?**).

First, our analytical results clearly show that though patients might incur an additional cost for technological support, many chronically ill patients will still benefit from the introduction of telemedicine in terms of access to care, as telemedicine increases both the geographical coverage and the capacity of specialist providers. Patients using telemedicine also enjoy a reduced travel burden and reduced dependence on others for travel. We also find instances where a lower fee for service will be optimal even for revenue-maximizing providers who implement telemedicine.

A second barrier is uncertainty about the differential effects of telemedicine. While some patients who have not been treated before (say, those who live far away or with serious motion disorders) will gain access to care, we do find that the benefit from telemedicine is not uniform for all patients, and indeed we show that some patients unexpectedly will be worse off with the introduction of telemedicine. For instance, after the introduction of telemedicine, those patients who live closer to the clinic will face busier providers, shorter visit times, and higher congestion costs. Hence it is important to recognize that not all specialist groups or patient populations will benefit equally. In addition, we have seen in our clinical studies that some of the patients who will be worse off might belong to the socio-economic strata that are highly desirable for the healthcare providers. Hence special care must be taken to not lose the goodwill of this population segment while expanding the geographic reach of the clinic. This brings about the policy option of introducing differential deployment of subsidies based on the degree of heterogeneity within the patient population.

A third barrier is the restrictive licensure laws in the United States, which require a practitioner to obtain a full license to deliver telemedicine care across state lines. Our results show that telemedicine has tremendous clinical and economic benefits, mainly for underserved populations. Hence several initiatives are already underway to relax these legal constraints (**?**).

The final barrier is the lack of clear guidelines for reimbursement of the specialist (**?**). Our results suggest that both specialist physicians and patients will likely benefit from telemedicine, even without any external subsidies, so long as the specialists are fairly reimbursed for telemedicine visits, even if the new (uniform) fees per face-to-face or telemedicine visit are set lower than the current fees for face-to-face office visits. For the reasons shown above, political leadership should act to remove the various legal barriers with respect to the deployment of and reimbursement for using telemedicine technology while treating chronically ill patients. In fact, several state Medicaid programs have already moved forward in allowing for proper reimbursement of telemedicine services (**?**). Congress, through the Cures Act, signed into law on December 13, 2016, directed the Centers for Medicare and Medicaid Services (CMS) to further study the use of technology (including telehealth) for the delivery of healthcare. This study will help decision makers better understand the major operational trade-offs in implementing telemedicine for chronic care and should assist in the planning of future field studies that could address the existing clinical and technological gaps identified above.

Acknowledgment: We thank the department editor, the anonymous associate editor and the three anonymous reviewers for their feedback and for their extremely useful suggestions. Earlier versions of this paper were presented at HICSS, INFORMS and MSOM conferences, and at several research seminars at London Business School, Stanford University, Yale University, University College London, and the University of Toronto and we thank all the participants for their useful feedback. We are also grateful to Professor Ray Dorsey, MD, for his practical and medical inputs on our analytical models.

References

Afèche, P. 2013. Incentive-compatible revenue management in queueing systems: Optimal strategic delay. *Manufacturing & Service Operations Management* **15**(3) 423–443.

AMA (Australian Medical Association). 2015. Setting medical fees and billing practices.

Anand, K.S., M.F. Pac, S.K. Veeraraghavan. 2011. Quality-speed conundrum: Tradeoffs in customer intensive services. *Management Science* **57**(1) 40–56.

Anderson, G., R. Herbert, T. Zeffiro, N. Johnson. 2004. Chronic conditions: Making the case for ongoing care: September 2004 update. Tech. rep., Partnership for Solutions.

Bavafa, H., L. M. Hitt, C. Terwiesch. 2017a. The impact of e-visits on visit frequencies and patient health: Evidence from primary care (July 28, 2017). *Working Paper* URL `http://dx.doi.org/10.2139/ssrn.2363705`.

Bavafa, H., S. Savin, C. Terwiesch. 2017b. Redesigning primary care delivery: Customized office revisit intervals and e-visits (May 20, 2017). *Working Paper* URL `http://dx.doi.org/10.2139/ssrn.2363685`.

Beck, C.A., D.B. Beran, K.M. Biglan, C.M. Boyd, E.R. Dorsey, P.N. Schmidt, R. Simone, A.W. Willis, N.B. Galifianakis, M. Katz, C.M. Tanner. 2017. National randomized controlled trial of virtual house calls for parkinson disease. *Neurology* **89**(11) 1152–1161.

Brady, M. K., J. J. Cronin Jr. 2001. Some new thoughts on conceptualizing perceived service quality: A hierarchical approach. *The Journal of Marketing* **66**(3) 34–49.

Brill, S. 2013. Bitter pill: Why medical bills are killing us. *Time* April 04, 2013.

Chen, H., M. Frank. 2004. Monopoly pricing when customers queue. *IIE Transactions* **36(6)** 569–581.

Chen, I.J., A. Gupta, W. Rom. 1994. A study of price and quality in service operations. *International Journal of Service Industry Management* **5(2)** 23–33.

Chen, P. W. 2012. How much can patients learn in a 15-minute doctor visit? *The New York Times* .

Crane, M. 2014. Integrating telemedicine into your practice. *Medical Economics* July 24, 2014.

Dávalos, M.E, M.T French, A.E. Burdick, S.C. Simmons. 2009. Economic evaluation of telemedicine: Review of the literature and research guidelines for benefit-cost analysis. *Telemedicine J and e-Health* **15(10)** 933–948.

Debo, L.J., C. Parlour, U. Rajan. 2012. Signaling quality via queues. *Management Science* **58(5)** 876–891.

Dorsey, E.R., B.P. George, B. Leff, A.W. Willis. 2013a. The coming crisis: Obtaining care for the growing burden of neurodegenerative conditions. *Neurology* **80** 1989–96.

Dorsey, E.R., V. Venkataraman, M.J. Grana, M.T. Bull, B.P. George, C.M. Boyd, C.A. Beck, B. Rajan, A. Seidmann, K.M. Biglan. 2013b. Randomized, controlled trial of "virtual housecalls" for Parkinson disease. *The Journal of the American Medical Association (JAMA) Neurology* **70(5)** 565–70.

Dussault, G., M. C. Franceschini. 2006. Not enough there, too many here: Understanding geographical imbalances in the distribution of the health workforce. *Human Resources for Health* **4(12)**.

Edelson, N.M., D.K. Hilderbrand. 1975. Congestion tolls for Poisson queuing processes. *Econometrica: Journal of the Econometric Society* **43(1)** 81–92.

Edwards, W. 2013. *Utility theories: Measurements and applications*, vol. 3. Springer Netherlands.

Farquhar, P.H. 1984. State of the art – Utility assessment methods. *Management Science* **11**(30) 1283–300.

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

Hargreaves, S. 2013. Cash-only doctors abandon the insurance system. *CNN Money* 11 Jun, 2013.

Hariri, S., K. J. Bozic, C. Lavernia, A. Prestipino, H. E. Rubash. 2007. Medicare physician reimbursement: past, present, and future. *The Journal of Bone and Joint Surgery* **89(11)** 2536–2546.

Hassin, R. 2016. *Rational Queueing*. Chapman & Hall/CRC Series in Operations Research, Taylor & Francis.

Hassin, R., M. Haviv. 2003. *To Queue or Not to Queue, Equilibrium Behaviour in Queuing Systems*. Kluwer Academic Publishers.

Hersh, W.R., D.H. Hickam, S.M. Severance, T.L. Dana, K.P. Krages, M. Helfand. 2006. Telemedicine for the Medicare population: Update. *AHRQuality* **131** 1–41.

Jauhar, S. 2014. Busy doctors, wasteful spending. *The New York Times* July 20 2014.

Jemal, A., E. Ward, Y. Hao, M. Thun. 2005. Trends in the leading causes of death in the United States, 1970-2002. *Journal of the American Medical Association* **294(10)** 1255–1259.

Khan, M. A., Y. Sun. 2002. Non-cooperative games with many players. R.J. Aumann, S. Hart, eds., *Handbook of Game Theory with Economic Applications*, *Handbook of Game Theory with Economic Applications*, vol. 3, chap. 46. Elsevier, 1761–1808.

Kong, M.C., F.T. Camacho, S.R. Feldman, R.T. Anderson, R. Balkrishnan. 2007. Correlates of patient satisfaction with physician visit: differences between elderly and non-elderly survey respondents. *Health and Quality of Life Outcomes* **5(62)**.

Kumar, A., G. de Lagasnerie, F. Maiorano, A. Forti. 2014. Pricing and competition in specialist medical services: An overview for South Africa. *OECD Health Working Papers* (70).

Kvedar, J., M. J. Coye, W. Everett. 2014. Connected health: A review of technologies and strategies to improve patient care with telemedicine and telehealth. *Health Affairs* **33(2)** 194–199.

Larsen, C. 1998. Investigating sensitivity and the impact of information on pricing decisions in an M/M/1/∞ queuing model. *International Journal of Production Economics* **56** 365–377.

Littlechild, S.C. 1975. Optimal arrival rate in a simple queuing system. *International Journal of Production Research* **12** 391–397.

Mendelson, H. 1985. Pricing services: Queuing effects. *Communications of the ACM* **28** 312–321.

Mendelson, H., S. Whang. 1990. Optimal incentive–Compatible priority pricing for the M/M/1 queue. *Operations Research* **38**(5) 870–883.

Mistry, H. 2012. Systematic review of studies of the cost-effectiveness of telemedicine and telecare. Changes in the economic evidence over twenty years. *Journal of Telemedicine and Telecare* **1**(18) 1–6.

Mossialos, E., D. Srivastava. 2008. Pharmaceutical policies in Finland. Challenges and opportunities. *Observatory Studies Series* **10**.

Naor, P. 1969. The regulation of queue size by levying tolls. *Econometrica: Journal of the Econometric Society* **1** 15–24.

Pennic, J. 2014. Should you become a telemedicine physician? *HIT Consultant* 5 December, 2014.

Polisena, J., D. Coyle, K. Coyle, S. McGill. 2009. Home telehealth for chronic disease management: A systematic review and an analysis of economic evaluations. *International Journal of Technology Assessment in Health Care* **25(3)** 339–349.

Rajan, B., A. Seidmann, T. Tezcan. 2014. Service model with different service rates and prices. Tech. rep., University of Rochester. URL `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2584212`.

Rath, K. P. 1992. A direct proof of the existence of pure strategy equilibria in games with a continuum of players. *Economic Theory* **2**(3) 427–433.

Samii, A., P. Ryan-Dykes, R.A. Tsukuda, C. Zink, R. Franks, W.P. Nichol. 2006. Telemedicine for the delivery of health care in Parkinson's disease. *Journal of Telemedicine and Telecare* **12(1)** 16–18.

Schmeidler, D. 1973. Equilibrium points of nonatomic games. *Journal of Statistical Physics* **7**(4) 295–300.

Stone, A. 2012. Why waiting is torture. *The New York Times* August 18, 2012.

Sullivan, P. 2012. Dealing with doctors who take only cash. *The New York Times* 23 Nov, 2012.

Sundaram, R.K. 2007. *A First Course in Optimization Theory*. Cambridge University Press.

Thomas, L., G. Capistrant. 2017a. State telemedicine gaps analysis: Coverage and reimbursement. Tech. rep., American Telemedicine Association.

Thomas, L., G. Capistrant. 2017b. State telemedicine gaps analysis: Physician practice standards and licensure. Tech. rep., American Telemedicine Association.

Weinstein, R. S., A. M. Lopez, B. A. Joseph, K. A. Erps, M. Holcomb, G. P. Barker, E. A. Krupinski. 2014. Telemedicine, telehealth, and mobile health applications that work: Opportunities and barriers. *The American Journal of Medicine* **127(3)** 183–187.

Xu, Y., M. Armony, A. Ghose. 2016. The effect of online reviews on physician demand: A structural model of patient choice. Tech. rep., NYU.

Zanaboni, P., R. Wootton. 2012. Adoption of telemedicine: From pilot stage to routine delivery. *BMC Medical Informatics and Decision Making* **12**(1).

Ziller, E.C. 2003. *Health insurance coverage in rural America: A Chartbook*. University of Southern Maine, Muskie School of Policy Service, Institute of Health Policy.

Appendices

A. Example Case

We present below a motivating example of a chronically ill patient (name and details changed to protect her privacy). It describes one of many similar patient scenarios that formed the basis of our research model.

Ms. Janice Taylor is a 61-year-old female who lives alone in Utica, New York. She was first diagnosed with Parkinson's disease (PD) back in 2009. Two years later, she lost her job at a small local dry-cleaning store due to her worsening medical condition. After waiting for over seven months in 2012, she was finally seen by Dr. Reginald Dolchini, a PD specialist at the University of Rochester Medical Center. The trip to this doctor's office is about 140 miles, and it takes her driver some 2.5 hours each way in a good weather. She goes all the way to Rochester since she does not have any PD specialist in her immediate neighborhood. Dr. Dolchini is the closest academic specialist who was able to accept her and gain her trust.

Addressing the multiple symptoms related to PD, Janice takes on average 7 to 10 different medications (including Carbidopa-levodopa, Sinemet, Amantadine, Entacapone, Pramipexole, Effexor, and Atropair). These drugs treat the primary disease, as well as the more severe side effects of her PD medications. These side effects include, in her case, insomnia, constipation, drooling, and depression. Typically, the medications and dosages change as the disease progresses and the medical conditions of the patient worsen over time, or when some of the side effects of these medications get acute. This is one of the primary reasons that PD patients need to be seen periodically by their specialist.

Her specialist prefers to see most of his patients on average four times a year in order to continuously monitor their condition, adjust their dosages, and recommend other complementary tests or treatments. We noticed that historically the actual annual count of past patient visits ranges between two and six. Visits are added due to unexpected disease progression, severe complications, or other nontrivial comorbidities. This happened to Janice about three years ago: after a new medication was introduced, she started suffering from symptoms resembling arrhythmia, but she had to wait days before she could get an appointment with Dr Dolchini, who was away at a conference and already had a long list of patients waiting to see him in the week of his return. Overall, Janice made a total of six visits to the clinic in 2013. A year later, Janice fell off the stairs while visiting her daughter; she ended up being hospitalized for several weeks. Upon her discharge, she needed to wait for a few more weeks before she could check with her specialist. In total, she visited Dr. Dolchini

only three times in that year.

The movement disorder clinic is very busy and hence they schedule only one patient visit at a time. They know well that the specialist's schedule and patients' demands are highly unpredictable beyond the next few months. In order to maintain sufficient capacity for any urgent patient needs—and to provide a reasonable service level for his current cases—Dr. Dolchini limits his panel size to only 200 active patients (as compared to 2,000-2,500 patients for a primary care physician). This is typical for his other academic colleagues at that group. Every time Janice visits the clinic in person, she uses her credit card to pay the typical follow-up visit fee of $ 250 at the front desk. She then submits the bill to her (private) insurance company for a partial reimbursement, as the clinic is out of the locally approved insurance network. Moreover, most of the leading insurance networks still do not cover telemedicine services for PD patients. These patients and nursing homes now directly pay the University PD clinic about $ 150 per each virtual visit.

B. List of Notations

x	-	Patient index, the distance of the patient from the specialist
X_m	-	Upper bound on the patient index, the maximum distance from the specialist
m	-	Reward per visit to the specialist; subscripts "i" and "t" are for in-person and telemedicine visits respectively
M_v	-	Upper bound on $m - t(x)$; that is, $m - t(0)$
f, F	-	pdf and cdf of the distance between patients and the specialist
Λ	-	Total arrival rate if the specialist decides to serve all patients
p	-	Price charged by a revenue-maximizing specialist
μ	-	Average service rate of a revenue-maximizing specialist
Λ	-	Total arrival rate if the specialist decides to serve all the patients
λ	-	Arrival rate of patient visits at the specialist
β	-	Co-insurance rate (a fraction between 0 and 1)
$Q(\mu)$	-	Cost or disutility to the patient as a function of the service rate
$\mathbb{E}W$	-	Expected time spent by the patient in the system
c	-	Opportunity cost per unit time for patients
$\mathcal{A} = \{i, t, o\}$	-	Action set for the patients; "i" denotes in-person treatment, "t" denotes telemedicine treatment, and "o" denotes no treatment
g_x	-	Distribution on $\mathcal{A}$ for a patient indexed by x
$t(x)$	-	Transportation cost (perceived) for the patient located at distance x from the specialist
$R(p, \mu)$	-	Revenue function for the revenue-maximizing specialist who does not offer telemedicine, defined in (**??**)
$U(\lambda, \mu)$	-	Total utility function for the welfare-maximizing specialist, defined in (**??**)

s	-	Amortized cost to the patient for setting up telemedicine visits
α	-	Proportion of visits possible via telemedicine; for a $(1-\alpha)$ fraction of the total visits, the patient has to visit the hospital or specialist in person (including emergency visits)
λ_{in}	-	Arrival rate of in-person mode patients at the specialist
λ_{tm}	-	Arrival rate of telemedicine mode patients at the specialist
λ_i	-	Arrival rate of in-person visits at the specialist
λ_t	-	Arrival rate of telemedicine visits at the specialist
$D(p,\mu)$	-	Location of the patient who is indifferent between choosing treatment and not seeking treatment when the telemedicine mode is not available
x_{ID}	-	Location of the patient who is indifferent between the two modes of treatment
x_{TM}	-	Location of the patient who is indifferent between choosing the telemedicine mode and not seeking treatment when the telemedicine option is available
$\bar{R}(p,\mu)$	-	Revenue function for the revenue-maximizing specialist who offers telemedicine, defined in (**??**)
δ	-	Proportionality constant for the service quality function
t_0	-	Transportation cost per unit distance for the travel burden function

C. Technical assumptions

In order to obtain the optimal decisions for the revenue-maximizing specialist we need to make certain assumptions. We list these assumptions in this section and explain why they are plausible. Throughout we assume that the derivative of a function denotes its right derivative at the left boundary and its left derivative at the right boundary of its domain. First, we assume that $m < \infty$ and $Q(\mu) < \infty$ for all $\mu \geq 0$. We make the following technical assumptions on F, t, and Q.

Assumptions on F and t:

A.1 $F : [0, X_m] \to [0,1]$ and $t : [0, X_m] \to \mathbb{R}$ are twice differentiable, and $F' > 0$, $F'' \leq 0$ on $[0, X_m]$.

A.2 $t : [0, X_m] \to \mathbb{R}$ is convex and strictly increasing on $[0, X_m]$.

The assumption that t is a convex increasing function (such as $t(x) = x^2$)—that is, $t'(x) > 0$ and $t''(x) \geq 0$—implies that as distance increases, the travel burden increases at a nondecreasing rate. The assumption is true in many general cases in which patients dealing with chronic conditions are old and therefore relatively less mobile. Many of these patients are also physically handicapped and need help from caretakers in case of travel. The burden thus is likely to increase at an increasing rate with distance. One can also expect $f'(x) < 0$, as the population density decreases with distance from the specialist (such as an exponential distribution, $f(x) = e^{-x}$), who is typically located in an urban area, as the population density tends to be lower in rural areas.

For the analysis in Section **??** the definition of X_m has to be modified slightly. Let M denote the maximum distance of a patient who might seek treatment from the specialist, where $M =$

$t^{-1}(\alpha/(1-\alpha)m_t + m_i) \geq t^{-1}(m_i)$; α is the fraction of clinical visits possible via telemedicine, and the remaining fraction, $(1-\alpha)$, involves visiting the specialist in person. We again use f and F to denote the pdf and the cdf for the distribution of the patients on $[0, M]$, respectively. We assume that Assumption **??** holds for $X_m = M$.

Assumptions on Q:

B.1 $Q : \mathbb{R}_+ \to \mathbb{R}_+$ is a strictly increasing function.

B.2 $Q : \mathbb{R}_+ \to \mathbb{R}_+$ is a convex function with continuous first and second derivatives.

Based on our discussion in the introduction, we assume that a patient's utility increases with time spent with the specialist and thus decreases with the service rate chosen by the specialist. Assumption **??**, which combined with Assumption **??** gives $Q'(\mu) > 0$, for $\mu \geq 0$, helps in modeling the higher cost to the patient (and thus the lower value to the patient) as the service rate increases. Assumption **??** also helps model the fact that the specialist is heavily penalized for working too fast. In addition, the convexity in Assumption **??** helps us simplify sufficient conditions for finding the optimal values. Assumptions **??** and **??** also imply that

$$\lim_{\mu \to \infty} Q(\mu) = \infty. \tag{19}$$

This ensures that if the specialist spends very little time with the patients, the patients will get no utility from the service. In addition, Assumption **??** implies $Q'(\mu) < \infty$, $Q''(\mu) < \infty$, for all $\mu \in \mathbb{R}_+$, which we use in the proofs of our main results.

D. Non-cooperative non-atomic games

Here we provide a general description of non-cooperative non-atomic games, which we use to study patient choices. We use the terminology introduced here in Section **??**. In the models we study, each patient has a choice between different treatment modes (including no treatment) and receives a reward based on his choice, his intrinsic value from treatment, the specialist's actions, and other patients' choices. We next define a general form for a game to capture these features.

We follow the terminology in **?**. Consider a continuum of patients (or players) indexed by $x \in [0, T]$ (for example, x denotes their utility from treatment), for some $T > 0$ and its Borel σ-algebra, endowed with an absolutely continuous probability measure μ with respect to the Lebesgue measure. Measure μ is used to assess how patients are distributed in this interval. Assume that each patient has to choose one of n treatment modes (or activities), and let $\mathcal{A} = \{1, \ldots, n\}$ denote the action set for patients. For our purposes it is enough to consider a discrete $\mathcal{A}$ and an action set that is not dependent on x. Denote

$$\mathcal{P} = \left\{ x = (x_1, \ldots, x_n) \in \mathbb{R}^n | x_i \geq 0, i = 1, 2, \ldots, n, \sum_{i=1}^{n} x_i = 1 \right\}. \tag{20}$$

Let $g_x \in \mathcal{P}$ denote a distribution on $\mathcal{A}$ for $x \in \mathbb{R}^+$, and let $G = \{g_x : x \geq 0\}$. Let $G^{-1}(x)$ denote the actions of all patients, excluding those with index x, and let $\Psi_a(\Gamma, G^{-1}(x), x)$ denote the reward received by a patient with index x if he chooses action a, all the patients follow the strategy profile G, and all the other external parameters are captured by Γ. (In our context Γ is used to denote the actions chosen by the specialist.) A T-strategy is a measurable function $\hat{x}$ from $[0,T]$ to P. Therefore, for $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_n)$, $\hat{x}_i$ is μ-integrable. We restrict our attention to T-strategies as in **?**.

A T-strategy G is said to yield a Nash equilibrium, for a given Γ, if it satisfies

$$\sum_{a \in \mathcal{A}} g_x(a) \Psi_a(\Gamma, G^{-1}(x), x) \geq \sum_{a \in \mathcal{A}} p_x(a) \Psi_a(\Gamma, G^{-1}(x), x) \tag{21}$$

μ-a.s. for $x \in \mathbb{R}^+$ and for any $p_x \in \mathcal{P}$. Inequality in (**??**) implies that under the strategy profile G no patients are better off by deviating from their choices in terms of expected utility. From here on we use "equilibrium" to refer to a Nash equilibrium.

In our setting, the formulation of the non-atomic games is slightly different from the extant literature (**???**). In **?**, for example, the players are indexed by a finite closed interval $[0,T]$, and this interval is mainly used for indexing purposes only. In our setting the index of a patient is associated with the utility a patient gets from receiving treatment, for example, depending on the distance $t \in [0,T]$ of the patient from the specialist. Unlike in **??**, and **?**, there might be multiple patients with the same index. Therefore, it is not clear how g_x should be interpreted. However, we show that in our setting, we can find a pure equilibrium strategy where patients with the same index follow the same strategy (in the a.s. sense). In general, one can consider g_x for a mixed strategy to be the distribution of patients with index x on how to choose each available action. In addition, because we assume that μ is an absolutely continuous probability measure with respect to the Lebesgue measure and because of our cost function, if a group of players with total measure zero change their actions, the payoffs to other players do not change. Hence the interpretation of g_x is not crucial for the applications we focus on.

E. General functions for disutility due to congestion

Our results hold under a more general function, denoted by θ, for the disutility due to congestion. Specifically, the following set of conditions on θ is sufficient for our results to hold.

- C.1 For $\mu > 0$, $\theta(\cdot, \mu) : \mathbb{R}_+ \to \mathbb{R}_+$ is a twice-differentiable, strictly increasing convex function on $[0, \mu)$, $\lim_{\lambda \uparrow \mu} \theta(\lambda, \mu) = \infty$.
- C.2 For $\lambda > 0$, $\theta(\lambda, \cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ is a twice-differentiable, strictly decreasing convex function on (λ, ∞) with $\lim_{\mu \to \infty} \theta(\lambda, \mu) = 0$, $\lim_{\mu \downarrow \lambda} \theta_\mu(\lambda, \mu) = -\infty$, and $\lim_{\mu \to \infty} \theta_\mu(\lambda, \mu) = 0$.

C.3 Given $M > 0$, there exists $\epsilon_M > 0$ such that $\theta(\lambda, \lambda + x) > M$ and $\theta_\mu(\lambda, \lambda + x) < -M$ for any $x \in (0, \epsilon_M)$.

For example, our results are valid under the following class of functions:

$$\theta(\lambda, \mu) = \sum_{i=1}^{I} \frac{a_i}{(\mu - \lambda)^i} \tag{22}$$

for a finite $I \geq 1$ and $a_i \geq 0$ for all $i \in [1, \ldots, I]$ and $a_k > 0$ for at least one $k \in [1, \ldots, I]$. The case $I = 1$ reduces to the linear case that we primarily focus on.

F. Proofs of the results in Section ??

We first characterize the optimal actions for revenue-maximizing and welfare-maximizing specialists in Appendix **??** and Appendix **??**, respectively. Then we use these results to prove Proposition **??** and Theorem **??** in Appendix **??**.

We start with the following result, which establishes some structural properties for V.

LEMMA 1. *If F and t are twice differentiable, t is convex, strictly increasing, and $F' > 0$, $F'' \leq 0$ on $[0, X_m]$, then we have the following:*

i. $V : [0, \Lambda] \to \mathbb{R}$ is a thrice-differentiable function on $[0, \Lambda]$.

ii. $V' : [0, \Lambda] \to \mathbb{R}$ is a decreasing concave function. Thus, we have $V'''(\lambda) \leq 0$ for all $\lambda \in (0, \Lambda)$.

*Proof of Lemma **??**:* By Assumptions **??** and **??**, F and t are twice differentiable, t is convex strictly increasing, and $F'(x) > 0$, $F''(x) \leq 0$ for all $x \in [0, X_m]$. Without loss of generality, we assume that $F(X_m) = 1$, as otherwise we can redefine it by conditioning on the fact all patients seeking treatment should be closer than X_m. Let $H(u) = F^{-1}(u)$ for $u \in [0, 1]$. Since F is strictly increasing, it is also invertible. Then, by our assumption that $F'(x) > 0$,

$$H'(u) = \frac{1}{F'(H(u))} \qquad \text{and} \qquad H''(u) = -\frac{F''(H(u))H'(u)}{(F'(H(u)))^2}.$$

for all $u \in [0, 1]$. Therefore,

$$H'(u) > 0 \text{ and } H''(u) \geq 0 \qquad \forall u \in [0, 1]. \tag{23}$$

If the arrival rate is λ, then the location of the farthest patient who seeks treatment is given by $H(\lambda/\Lambda)$. By Leibniz's rule, $V'(\lambda) = m - t(F^{-1}(\lambda/\Lambda))$ for $\lambda \in (0, \Lambda)$, so

$$V'(\lambda) = m - t\left(H(\lambda/\Lambda)\right), V''(\lambda) = -\frac{1}{\Lambda} t'\left(H(\lambda/\Lambda)\right) H'(\lambda/\Lambda), \text{ and} \tag{24}$$

$V'''(\lambda) = -1/\Lambda^2 \left[(H'(\lambda/\Lambda))^2 t''\left(H(\lambda/\Lambda)\right) + t'\left(H(\lambda/\Lambda)\right) H''(\lambda/\Lambda)\right]$. We are now ready to prove the lemma.

A1) Note that $V(\lambda)$ is a thrice-differentiable function, since t and F are twice differentiable as well.

A2) We need to show that $V'(\lambda)$ is a decreasing concave function. By (**??**) and the assumption $t'(x) > 0$, $V''(\lambda) \leq 0$; by the fact that $t''(x) \geq 0$, and by (**??**), $V'''(\lambda) \leq 0$ on $\lambda \in [0, \Lambda F^{-1}(M)]$. $\square$

F.1. Optimal decisions of a revenue-maximizing specialist

We next find the optimal decisions of a revenue-maximizing specialist, stated in Lemma **??**, given patient utility function Ψ, defined in (**??**). The main technical difficulty is that in the definition of x^* in (**??**) the stability constraint makes the analysis around this boundary very challenging. We tackle this problem by showing that $\mu \geq \lambda(p, \mu) + \delta$ for some $\delta > 0$, if the revenue is positive.

The optimization problem in (**??**) has price and service rates as decision variables for the specialist. The equilibrium arrival rate is then determined by the price and service rate set by the specialist, as given by (**??**). It is easier to solve the optimization problem with the arrival rate and the service rate as decision variables instead. Because the arrival rate is automatically determined by the specialist's choice of service rate and price, this new optimization problem is equivalent to the original one, (**??**), as we show next.

Preliminaries: Let $p : [0, \Lambda] \times \mathbb{R}_+ \to \mathbb{R}_+$ be defined by

$$p(\lambda, \mu) = \inf \left\{ p \in \mathbb{R} : m - t(F^{-1}(\lambda/\Lambda)) - Q(\mu) - \beta p - \frac{c}{\mu - \lambda} \leq 0 \right\} \vee 0, \tag{25}$$

where we take $F^{-1}(0) = 0$ and $F^{-1}(1) = X_m$. Hence $p(\lambda, \mu)$ can be interpreted as the equilibrium price given the service rate μ and the arrival rate λ. In other words, given the arrival rate and the service rate, the price is chosen such that the marginal patient (with threshold distance $x^* = F^{-1}(\lambda/\Lambda)$) is indifferent between seeking treatment and not seeking treatment. (Otherwise, the price can simply be increased for higher revenue.) We need to ensure that the price is non-negative for practical reasons. We define $\tilde{R} : [0, \Lambda] \times \mathbb{R} \to \mathbb{R}$ as follows:

$$\tilde{R}(\lambda, \mu) = \begin{cases} p(\lambda, \mu)\lambda, & \text{if } \lambda \in [0, \Lambda] \text{ and } \mu > \lambda \\ 0, & \text{otherwise} \end{cases}. \tag{26}$$

The specialist's objective is to maximize his revenue by choosing an appropriate service rate μ and arrival rate λ. We will thus be able to define the following optimization problem:

$$\tilde{R}^* = \sup_{\lambda \geq 0,\ \mu \geq 0} \tilde{R}(\lambda, \mu). \tag{27}$$

Before we verify that the optimization problem in (**??**) is equivalent to (**??**), we first show in the following lemma that the specialist can only choose from a range of service rates for the revenue to be non-zero. Let $\kappa = \Lambda F(t^{-1}(m - Q(c/M_v)))$.

LEMMA 2. *If $\mu \geq Q^{-1}(M_v)$ or if $\mu \leq \lambda + c/M_v$, then $\tilde{R}(\lambda, \mu) = 0$ for any $\mu \geq 0$ and $\lambda \in [0, \Lambda]$. Also, if $\lambda \geq \kappa$, then $\tilde{R}(\lambda, \mu) = 0$ for all $\mu \geq 0$.*

Proof of Lemma **??***:* Let $\mu \geq Q^{-1}(M_v)$. We note that the inverse exists by Assumptions **??** and **??**, and by (**??**). Then $Q(\mu) \geq M_v$. Let μ be such that $\mu > \lambda$. If not, $\tilde{R}(\lambda, \mu) = 0$ from (**??**). Since $c > 0$ and $\mu > \lambda$, we have $Q(\mu) + c/(\mu - \lambda) > M_v$, which implies, from (**??**), that $x^* = 0$, and so $p(\lambda, \mu) = 0$ from (**??**). Hence $\tilde{R}(\lambda, \mu) = 0$ from (**??**). Now we find the lower limit for μ. Let $\mu \leq \lambda + c/M_v$. Rearranging terms, we have, and since $Q(\mu) \geq 0$, $Q(\mu) + c/(\mu - \lambda) \geq M_v$, $x^* = 0$ from (**??**), and so $p(\lambda, \mu) = 0$ from (**??**).

Finally, if $\lambda \geq \kappa$, then $m - t(F^{-1}(\lambda/\Lambda)) \leq Q(c/M_v)$.

If $\mu \leq \lambda + c/M_v$, then $\tilde{R}(\lambda, \mu) = 0$ by the first part. If $\mu > \lambda + c/M_v$, then $Q(\mu) > Q(c/M_v)$. This, combined with $m - t(F^{-1}(\lambda/\Lambda)) \leq Q(c/M_v)$, gives $m - t(F^{-1}(\lambda/\Lambda)) - Q(\mu) \leq 0$. Thus $p(\lambda, \mu) = 0$ by (**??**), giving the desired result. □

Throughout, we assume that $Q(c/M_v) < M_v$, as otherwise $p(\lambda, \mu) = 0$ for all $\mu > \lambda \geq 0$ by the first part of Lemma **??** and (**??**). Hence $\kappa < \Lambda$. Lemma **??** follows from the fact that the specialist can work neither too quickly nor too slowly. If he treats patients too quickly, then the cost of quality will be too high, because $Q(\mu)$ is decreasing in μ, resulting in a negative net utility for all the patients. Similarly, if the specialist treats patients too slowly, congestion, and hence waiting cost, will rise, resulting in a negative net utility for all the patients. (We note that Lemma **??** is similar to the quantities $A_1(\alpha)$ and $A_2(\alpha)$ on page 43 of **?**, except that our result is valid for a more general model.)

Mathematically, Lemma **??** helps in finding lower and upper bounds for the optimal service rate. We will thus be able to write the optimization problem in (**??**) using Lemma **??** as

$$\tilde{R}^* = \sup_{\lambda \geq 0,\ \lambda + \frac{c}{M_v} < \mu \leq Q^{-1}(M_v)} \tilde{R}(\lambda, \mu). \tag{28}$$

The boundary conditions in (**??**) follow from Lemma **??**. We have the following lemma to prove that the optimization problem in (**??**) is equivalent to (**??**):

LEMMA 3. *For R defined as in (**??**) and $\tilde{R}$ as in (**??**), if $R(p, \mu) > 0$ for $p \geq 0$ and $\mu \geq 0$, then $R(p, \mu) = \tilde{R}(\lambda(p, \mu), \mu)$, and if $\tilde{R}(\lambda, \mu) > 0$ for $\lambda \geq 0$ and $\mu \geq 0$, then $\tilde{R}(\lambda, \mu) = R(p(\lambda, \mu), \mu)$. Hence $\tilde{R}^* = R^*$.*

Proof of Lemma **??***:* Let $\mu' > 0$ and $p' > 0$ be such that $R(p', \mu') > 0$ (if $\mu' = 0$ or $p' = 0$, then $R(p', \mu') = 0$ by (**??**), (**??**), and (**??**)). If $R(p', \mu') > 0$, then $\lambda(p', \mu') > 0$, and so $x^* > 0$ by (**??**). In

addition, $x^* < X_m$ by (**??**). From Assumptions **??** and **??**, $F^{-1}(t(.))$ is continuous on $(0,1)$ also. This implies by (**??**) and (**??**) that

$$m - t(F^{-1}(\lambda(p',\mu')/\Lambda)) - Q(\mu') - \beta p' - \frac{c}{\mu - \lambda(p',\mu')} = 0$$

and $\mu' > \lambda(p',\mu')$. Thus, by (**??**), $p(\lambda(p',\mu'),\mu') = p'$. Hence $\tilde{R}(\lambda(p',\mu'),\mu') = R(p',\mu')$.

Now let $\lambda' > 0$ and $\mu' > 0$ be such that $\tilde{R}(\lambda',\mu') > 0$ (if $\lambda' = 0$ or $\mu' = 0$, then $\tilde{R}(\lambda',\mu') = 0$ by (**??**)). This implies by (**??**) that $p(\lambda',\mu') > 0$ and $\lambda' < \mu'$. Hence, by Lemma **??**, $\lambda' < \Lambda$, and by (**??**),

$$m - t(F^{-1}(\lambda'/\Lambda)) - Q(\mu') - \beta p(\lambda',\mu') - \frac{c}{\mu' - \lambda'} = 0.$$

By (**??**), $\lambda(p(\lambda',\mu'),\mu') = \lambda'$. Hence $\tilde{R}(\lambda',\mu') = R(\lambda(p(\lambda',\mu'),\mu')$. The fact that $\tilde{R}^* = R^*$ follows from Lemma **??**. □

We can now concentrate on finding the optimal service rate for the specialist and the optimal λ (the arrival rate) that the specialist must choose. The optimal price can then be determined using (**??**).

Solution to the optimization problem: We solve the optimization problem in (**??**) in two sequential steps. First, we solve the following optimization problem for fixed $\lambda \geq 0$:

$$\tilde{R}^*(\lambda) = \sup_{\lambda + \frac{c}{M_v} \leq \mu} \tilde{R}(\lambda,\mu). \tag{29}$$

That is, we find the optimal service rate for a given arrival rate. Let $\gamma : \mathbb{R}^+ \to \mathbb{R}$ be defined as follows:

$$\gamma(\lambda) = \left\{ \mu : Q'(\mu) - \frac{c}{(\mu-\lambda)^2} = 0 \text{ and } \mu > \lambda \right\}. \tag{30}$$

We start with the following technical result.

LEMMA 4. *Mapping $\gamma : (0,M) \to \mathbb{R}$ is a well-defined continuous function for any finite constant $M > 0$. Also, $0 < \gamma'(\lambda) = 2c/(2c + (\gamma(\lambda) - \lambda)^3 Q''(\gamma(\lambda))) \leq 1$.*

Proof of Lemma **??***:* Let $M > 0$ denote a finite constant. Define

$$J(\lambda,\mu) = Q'(\mu) - \frac{c}{(\mu-\lambda)^2}.$$

By Assumptions **??** and **??**, Q' is nondecreasing and continuous. Also, because $c/(\mu-\lambda)^2$ is a decreasing continuous function for $\mu > \lambda$ with its values dense in $(0,\infty)$, and from Assumptions **??** and **??**, there exists $\mu_\lambda > \lambda$ such that $J(\lambda,\mu_\lambda) = 0$.

Because Q' is nondecreasing, $\mu_{\lambda_1} < \mu_{\lambda_2}$, if $\lambda_1 < \lambda_2$. Therefore,

$$\mu_\lambda - \lambda \geq \sqrt{c/Q'(\mu_0)}, \text{ for all } \lambda \in [0, M]. \tag{31}$$

Next we prove that there exists a continuously differentiable function $\gamma : (0, M) \to \mathbb{R}^+$ such that $\gamma(\lambda) > \lambda + \delta$ for some $\delta > 0$ and

$$J(\lambda, \gamma(\lambda)) = 0, \quad \text{for } \lambda \in (0, M). \tag{32}$$

Before we proceed with the proof of (**??**), we recall the implicit function theorem (IFT). The IFT states the following: If J is defined on an open disk containing (λ, μ), where $J(\lambda, \mu) = 0$, $J_\mu(\lambda, \mu) \neq 0$, and J_μ and J_λ are continuous on the disk, then the equation $J(\lambda, \mu) = 0$ defines μ as a function of λ; i.e., there exists a function γ such that $J(\lambda', \gamma(\lambda')) = 0$ in a neighborhood of (λ, μ). We now check whether all the conditions hold to allow us to apply the theorem to prove the existence of $\gamma(\lambda)$ for $\lambda \in (0, M)$.

By the discussion above, given $\lambda > 0$, there exists a unique $\mu_\lambda > \lambda$ such that $J(\lambda, \mu_\lambda) = 0$. Also, the partial derivatives of J with respect to λ and μ, J_λ and J_μ respectively, are continuous in the region $\lambda > 0$ and $\mu > \lambda$. In addition,

$$\frac{\partial J}{\partial \mu}(\lambda, \mu_\lambda) = Q''(\mu_\lambda) + \frac{2c}{(\mu_\lambda - \lambda)^3} = \frac{\partial^2 \tilde{R}}{\partial \mu^2}(\lambda, \mu_\lambda) > 0,$$

because of Assumption **??** and $\mu_\lambda > \lambda$. Hence all the conditions of the implicit function theorem hold, so for any $\lambda \in (0, M)$ there exists a unique function γ_λ and a neighborhood $\mathcal{A}_\lambda$ of λ such that $J(\lambda', \gamma_\lambda(\lambda')) = 0$ for all $\lambda' \in \mathcal{A}_\lambda$.

Although the IFT proves the existence of γ_λ, we still have not proved that there exists a unique differentiable function γ on $(0, M)$ and that $\gamma(\lambda) > \lambda$. We next prove that for $\lambda \neq \lambda'$, $\gamma_\lambda = \gamma_{\lambda'}$ on $\mathcal{A}_\lambda \cap \mathcal{A}_{\lambda'} (\neq \emptyset)$, proving uniqueness. The fact that $\gamma(\lambda) > \lambda + \delta$ for some $\delta > 0$ follows from (**??**).

First, each γ_λ is continuous and differentiable by the IFT. Assume that there exists $\lambda' \in \mathcal{A}_\lambda$ such that $\gamma_\lambda(\lambda') < \lambda'$ and $\lambda' < \lambda$. This implies by the continuity of γ_λ that there exists $\lambda'' \in (\lambda', \lambda)$ such that $\lambda'' = \gamma_\lambda(\lambda'')$. However, this is not possible, since $Q'(\lambda'')$ is bounded by Assumptions **??** and **??**. Hence no such λ' exists. We can prove the conclusion for $\lambda' > \lambda$ similarly. Hence for all $\lambda' \in \mathcal{A}_\lambda$, $\gamma_\lambda(\lambda') > \lambda'$. However, there is a unique $\gamma_\lambda(\lambda') > \lambda'$ that satisfies (**??**), so $\gamma_\lambda = \gamma_{\lambda'}$ on $\mathcal{A}_\lambda \cap \mathcal{A}_{\lambda'} (\neq \emptyset)$. Therefore there is a unique γ that satisfies the conditions of the result.

By the IFT we have

$$\frac{d\gamma(\lambda)}{d\lambda} = -\frac{\frac{\partial J}{\partial \lambda}}{\frac{\partial J}{\partial \mu}} = \frac{2c}{2c + (\gamma(\lambda) - \lambda)^3 Q''(\gamma(\lambda))}.$$

From Assumption **??** and the fact that $\gamma(\lambda) > \lambda + \delta$, $\gamma'(\lambda) > 0$ and $\gamma'(\lambda) \leq 1$ for all $\lambda \in (0, M)$. $\square$

Next we show that γ gives the optimal service rate.

LEMMA 5. *Given $\lambda \in (0,\kappa)$, if there exists $\mu(>0)$ such that $\tilde{R}(\lambda,\mu)>0$, then $\gamma(\lambda)$ gives the optimal solution for (**??**); that is, $\tilde{R}^*(\lambda)=\tilde{R}(\lambda,\gamma(\lambda))$. If $\lambda \geq \kappa$, $\lambda=0$, or $p(\lambda,\mu)=0$ for all $\mu \geq 0$, then $\tilde{R}^*(\lambda)=0$.*

*Proof of Lemma **??**:* Let $\lambda \in (0,\kappa)$ and assume that there exists $\mu>0$ such that $\tilde{R}(\lambda,\mu)>0$. Next we show that given λ, $\gamma(\lambda)$ gives the optimal solution for (**??**). Let

$$p^+(\lambda,\mu)=\frac{1}{\beta}\left(m-t(F^{-1}(\lambda/\Lambda))-Q(\mu)-\frac{c}{\mu-\lambda}\right) \quad \text{and} \tag{33}$$

$$\tilde{R}^+(\lambda,\mu)=\lambda p^+(\lambda,\mu)=\frac{1}{\beta}\left(m-t(F^{-1}(\lambda/\Lambda))-Q(\mu)-\frac{c}{\mu-\lambda}\right)\lambda, \tag{34}$$

for $\lambda \in [0,\Lambda]$ and $\mu>\lambda$.

Note that

$$p^+(\lambda,\mu) \leq p(\lambda,\mu) \text{ and } \tilde{R}^+(\lambda,\mu) \leq \tilde{R}(\lambda,\mu). \tag{35}$$

The first inequality in (**??**) follows from the definition of $p(\lambda,\mu)$ (see (**??**)), and the second inequality follows from (**??**). We note that if $\tilde{R}^+(\lambda,\mu)<0,\ \forall \mu>\lambda$, then $p^+(\lambda,\mu)<0,\ \forall \mu>\lambda$ by (**??**) and thus $p(\lambda,\mu)=0,\ \forall \mu>\lambda$, from (**??**). Then $\tilde{R}(\lambda,\mu)=0,\ \forall \mu>\lambda$, from (**??**). On the other hand, if $\tilde{R}^+(\lambda,\mu)>0$, then $p^+(\lambda,\mu)>0$, so $p^+(\lambda,\mu)=p(\lambda,\mu)$. Also, $p^+(\lambda,\mu)=p(\lambda,\mu)$ if $p(\lambda,\mu)>0$. Therefore,

$$\tilde{R}^+(\lambda,\mu)=\tilde{R}(\lambda,\mu), \text{ if } p^+(\lambda,\mu)>0 \text{ or } p(\lambda,\mu)>0. \tag{36}$$

Given $\lambda \in (0,\kappa)$, we next solve

$$\sup_{\mu \geq \lambda+\frac{c}{M_v}} \tilde{R}^+(\lambda,\mu)=\sup_{\mu \geq \lambda+\frac{c}{M_v}} \frac{1}{\beta}\left(m-t(F^{-1}(\lambda/\Lambda))-Q(\mu)-\frac{c}{\mu-\lambda}\right)\lambda.$$

We use KKT necessary conditions to obtain the optimal service rate. Given $\lambda>0$, since the constraint $\mu \geq \lambda+c/M_v$ is linear in μ, regularity conditions are satisfied, so if μ^*_λ is optimal then there exists σ such that the following KKT necessary conditions are satisfied (see Proposition 6.5, **?**):

$$\begin{aligned}
\frac{\partial \tilde{R}^+}{\partial \mu}(\lambda,\mu^*_\lambda) &= -\sigma, \\
\mu^*_\lambda &\geq \lambda+\frac{c}{M_v}, \\
\sigma &\geq 0, \quad \text{and} \\
\sigma\left(\mu^*_\lambda-\left(\lambda+\frac{c}{M_v}\right)\right) &= 0.
\end{aligned}$$

We then have the following two cases for non-negative σ:

Case 1, $\boldsymbol{\sigma = 0}$: This implies that μ_λ along with $\sigma = 0$ will satisfy the KKT conditions if $\mu_\lambda \geq \lambda + c/M_v$. Also, from (**??**), $\partial \tilde{R}^+/\partial \mu = -J(\lambda, \mu)$, Hence $\mu_\lambda = \gamma(\lambda)$ is the unique solution.

Case 2, $\boldsymbol{\sigma > 0}$: This implies that $\mu^*_\lambda = (\lambda + c/M_v)$ and $\sigma = -\partial \tilde{R}^+/\partial \mu(\lambda, \mu^*_\lambda)$ is the only solution that will satisfy the KKT conditions if $\sigma = -\partial \tilde{R}^+/\partial \mu(\lambda, \mu^*_\lambda) > 0$.

Next we argue that we can ignore the solution given by Case 2 for the optimization problem (**??**). Note that if $\mu^*_\lambda = (\lambda + c/M_v)$ (i.e., Case 2 holds), from Lemmas **??** and **??**, we have $\tilde{R}(\lambda, \mu^*_\lambda) = 0$. Hence, $\tilde{R}^+(\lambda, \mu) \leq 0,\ \forall \mu > \lambda$, which implies $\tilde{R}^*(\lambda) = 0$ from the discussion above. We can take μ^*_λ as in Case 1, since $\tilde{R}(\lambda, \mu) \geq 0$. Otherwise, if $\tilde{R}^+(\lambda, \mu^*_\lambda) > 0$, then $\mu = (\lambda + c/M_v)$ cannot be optimal, so we need to consider only Case 1. Therefore, from (**??**), we obtain the result stated in the first part of the lemma.

Next we prove the last part of the lemma. If $p(\lambda, \mu) = 0,\ \forall \mu$, then $\tilde{R}(\lambda, \mu) = 0,\ \forall \mu$. If $\lambda \geq \kappa$, the result follows from Lemma **??**. $\square$

REMARK 3. By (**??**) and Lemmas **??** and **??**, the optimal service rate is increasing in the arrival rate. More interestingly, the derivative, γ', is bounded by 1. In other words, in response to an increase in the arrival rate, the optimal capacity increases at most by the same rate. The implication is that because the optimal service rate will not increase at a faster rate, utilization and congestion costs will not decrease (will increase if Q is strictly convex) with an increase in the arrival rate. Thus the average congestion cost for all patients will not decrease (will increase if Q is strictly convex).

We next solve the optimization problem for λ to complete the optimization problem in (**??**). We do so by substituting the optimal service rate for a given λ from Lemma **??**. Assume that $\tilde{\lambda} \in [0, \Lambda]$ satisfies the following equation (we show that the solution is unique if a solution exists):

$$V'(\tilde{\lambda}) + \tilde{\lambda} V''(\tilde{\lambda}) - Q(\gamma(\tilde{\lambda})) - \frac{c}{\gamma(\tilde{\lambda}) - \tilde{\lambda}} - \frac{c\tilde{\lambda}}{(\gamma(\tilde{\lambda}) - \tilde{\lambda})^2} = 0. \tag{37}$$

LEMMA 6. *If there exists (λ, μ) such that $\tilde{R}(\lambda, \mu) > 0$, then there exists a unique optimal λ^*; that is, $\tilde{R}(\lambda^*, \gamma(\lambda^*)) = \tilde{R}^* > 0$ and $\lambda^* = \tilde{\lambda}$. If no such (λ, μ) exists, $\tilde{R}^* = R^* = 0$.*

*Proof of Lemma **??**:* Assume that there exists $(\lambda, \mu) \geq 0$ such that $\mu > \lambda$, $\lambda \in [0, \Lambda]$, and $\tilde{R}(\lambda, \mu) > 0$. First note that $\mu > 0$ because $\mu > \lambda$ and $\lambda > 0$ by (**??**) because $\tilde{R}(\lambda, \mu) > 0$.

If $\tilde{R}(\lambda, \mu) > 0$ for some λ and μ satisfying the conditions above, and $\lambda^* \in [0, \Lambda]$ and $\mu^* > \lambda^*$ satisfies

$$\tilde{R}^+(\lambda^*, \mu^*) = \sup_{\lambda \in [0, \Lambda], \mu > \lambda} \tilde{R}^+(\lambda, \mu),$$

then by (**??**) and Lemma **??**, (λ^*, μ^*) must be the solution of (**??**) as well. We next show that such λ^* and μ^* exist.

We have

$$\sup_{\lambda\in[0,\Lambda],\mu>\lambda} \tilde{R}^+(\lambda,\mu) = \sup_{\lambda\in(0,\kappa],\mu>\lambda} \tilde{R}^+(\lambda,\mu) = \sup_{\lambda\in(0,\kappa]} \tilde{R}^+(\lambda,\gamma(\lambda)),$$

where the first equality follows from the condition $\tilde{R}(\lambda,\mu) > 0$ and Lemma **??** and the second inequality follows from (**??**) and Lemma **??**. By (**??**) and Lemma **??**, $\tilde{R}^+(\lambda,\gamma(\lambda))$ is continuous on $(0,\Lambda]$ and $\limsup_{\lambda\to 0} \tilde{R}^+(\lambda,\gamma(\lambda)) \le 0$, so there exists $0 < \epsilon < \kappa$ such that $\tilde{R}^+(x,\gamma(x)) < \tilde{R}^+(\lambda,\gamma(\lambda))$ for $0 < x \le \epsilon$. Hence

$$\sup_{\lambda\in(0,\kappa]} \tilde{R}^+(\lambda,\gamma(\lambda)) = \sup_{\lambda\in[\epsilon,\kappa]} \tilde{R}^+(\lambda,\gamma(\lambda)). \tag{38}$$

Also, since $V'(\lambda) = m - t(H(\lambda/\Lambda))$,

$$\beta\tilde{R}^+(\lambda,\gamma(\lambda)) = \left(V'(\lambda) - Q(\gamma(\lambda)) - \frac{c}{\gamma(\lambda)-\lambda}\right)\lambda$$

for $\lambda \in (0,\Lambda)$. By Lemma **??**, $\tilde{R}^+(\lambda,\gamma(\lambda))$ is continuous in λ on $[\epsilon,\kappa]$, so there exists $\lambda^* \in (0,\kappa]$ such that $\tilde{R}^+(\lambda^*,\gamma(\lambda^*)) = \sup_{\lambda\in[\epsilon,\kappa]} \tilde{R}^+(\lambda,\gamma(\lambda))$.

Using the second derivative of $\tilde{R}^+(\lambda,\gamma(\lambda))$, we next show that $\tilde{R}^+(\lambda,\gamma(\lambda))$ is concave in λ on $[\epsilon,\kappa]$. First, by (**??**),

$$\beta\frac{d\tilde{R}^+(\lambda,\gamma(\lambda))}{d\lambda} = V'(\lambda) - Q(\gamma(\lambda)) - \frac{c}{\gamma(\lambda)-\lambda} + \lambda V''(\lambda) - \frac{c\lambda}{(\gamma(\lambda)-\lambda)^2} \tag{39}$$

and

$$\beta\,\frac{d^2\tilde{R}^+(\lambda,\gamma(\lambda))}{d\lambda^2} = 2V''(\lambda) + \lambda V'''(\lambda) - \frac{2c}{(\gamma(\lambda)-\lambda)^2} - \frac{2c\lambda(1-\gamma'(\lambda))}{(\gamma(\lambda)-\lambda)^3}.$$

By Lemma **??**, Lemma **??**, Lemma **??**, and Assumptions **??** and **??**, each of the terms on the right-hand side is either negative or non-positive. Also, by Lemma **??**,

$$\beta\frac{d^2\tilde{R}^+(\lambda,\gamma(\lambda))}{d\lambda^2} < 0, \text{ for } \lambda \in (0,\kappa). \tag{40}$$

Thus, $\tilde{R}^+(\lambda,\gamma(\lambda))$ is concave in λ for $\lambda \in [\epsilon,\kappa]$. Recall that by our assumption there exists $\lambda \in [\epsilon,\kappa]$ such that $\tilde{R}^+(\lambda,\gamma(\lambda)) > 0$. Because $\tilde{R}^+(\lambda,\gamma(\lambda))$ is concave in λ and $\tilde{R}^+(\epsilon,\gamma(\epsilon)) < \tilde{R}^+(\lambda^*,\gamma(\lambda^*))$ and $\tilde{R}^+(\kappa,\gamma(\kappa)) < \tilde{R}^+(\lambda^*,\gamma(\lambda^*))$, for $\lambda \in [\epsilon,\kappa]$, we can use the necessary first-order condition for an optimal solution. The necessary first-order condition (using (**??**)) gives (**??**). Uniqueness is then ensured by strict concavity. Also, (**??**) has a solution, $\tilde{\lambda} \in [\epsilon,\kappa]$, because the optimal point is not at the boundaries. The second part of the lemma is obvious. □

Given functional forms for the utility functions, the optimal service rate can thus be determined using Lemma **??**. Finally, using (**??**), the price can be determined as well, solving the optimization problem in (**??**) by Lemma **??**.

F.2. Optimal decisions of a welfare-maximizing specialist

In Section **??**, we considered the specialist and the patients to be separate entities focused on their own self-interest. Here we consider the specialist and the patients as a single combined entity and solve the optimization problem in (**??**) for socially optimal decisions. The technical details are similar to those in Section **??**, so we do not repeat them here.

Assume that $\tilde{\lambda}_s \in [0, \Lambda]$ satisfies the following equation (we show that the solution is unique if a solution exists):

$$V'(\tilde{\lambda}_s) - Q(\gamma(\tilde{\lambda}_s)) - \frac{c}{\gamma(\tilde{\lambda}_s) - \tilde{\lambda}_s} - \frac{c\tilde{\lambda}_s}{(\gamma(\tilde{\lambda}_s) - \tilde{\lambda}_s)^2} = 0.$$

By (**??**) and since $V'(\lambda) = m - t(F^{-1}(\lambda/\Lambda))$ for $\lambda \in (0, \Lambda)$, the marginal change in the cumulative benefit function value for the welfare-maximizing specialist with an increase in the arrival rate is equal to the benefit received by the farthest patient ($m - t(x_\lambda)$) among the patients seeking treatment, a fact we use below.

LEMMA 7. *If there exists (λ, μ) such that $U(\lambda, \mu) > 0$, then there exists a unique optimal λ_s^*; that is, $U(\lambda_s^*, \gamma(\lambda_s^*)) = U^* > 0$. If $\tilde{\lambda}_s \in [0, \Lambda]$ exists, then $\lambda_s^* = \tilde{\lambda}_s$; otherwise, $\lambda_s^* = \Lambda$. If no such (λ, μ) exists, $U^* = 0$.*

Comparing Lemmas **??** and **??**, if the arrival rate per unit time at the revenue-maximizing specialist and the welfare-maximizing specialist is the same and if $U^*(\lambda) > 0$, then they both work at the same service rate. In other words, if the arrival rates of patients at the specialists are exogenous, then both revenue-maximizing and welfare-maximizing specialists spend equal amounts of time with the patients on average.

Before we prove Lemma **??**, we present a few preliminary results. The following result corresponds to Lemma **??** in this setting.

LEMMA 8. *If $\mu \geq Q^{-1}(M_v)$ or if $\mu \leq \lambda + c/M_v$, then $U(\lambda, \mu) \leq 0$.*

The proof is similar to that of Lemma **??** and thus is omitted. Using this result, the objective (**??**) can be written as $U^* = \sup_{\lambda \geq 0,\ \lambda + c/M_v \leq \mu \leq Q^{-1}(M_v)} U(\lambda, \mu)$, and, by (**??**),

$$U(0, \mu) = 0, \text{ for any } \mu \geq 0. \tag{41}$$

Next, we find the optimal service rate, μ, given the arrival rate, λ; that is, given $\lambda \geq 0$, we solve the optimization problem $U^*(\lambda) = \sup_{\lambda + c/M_v \leq \mu} U(\lambda, \mu)$. We then have the following lemma, which is similar to Lemma **??**, for $\gamma(\lambda)$ defined in (**??**).

LEMMA 9. *Given $\lambda \in (0, \Lambda]$, if there exists $\mu > \lambda$ such that $U(\lambda, \mu) > 0$, then $U^*(\lambda) = U(\lambda, \gamma(\lambda))$.*

Next we prove Lemma **??**.

Proof of Lemma **??***:* Assume that there exists (λ, μ) such that $U(\lambda, \mu) > 0$. We have

$$\sup_{0 \le \lambda \le \Lambda, \mu \ge 0} U(\lambda, \mu) = \sup_{0 < \lambda \le \Lambda, \mu \ge 0} U(\lambda, \mu) = \sup_{0 < \lambda \le Q^{-1}(M_v) \wedge \Lambda} U(\lambda, \gamma(\lambda)), \tag{42}$$

where the first equality follows from (**??**) and the second equality follows from the assumption that $U^* > 0$ and Lemma **??**.

Define

$$U^+(\lambda) = V(\lambda) - \left(Q(\gamma(\lambda)) + \frac{c}{\gamma(\lambda) - \lambda} \right) \lambda.$$

Because $U^*(\lambda_s^*) = \sup_{\lambda \ge 0} U^*(\lambda)$, and $U^* > 0$ by the assumption in the theorem,

$$\sup_{0 < \lambda \le Q^{-1}(M_v) \wedge \Lambda} U(\lambda, \gamma(\lambda)) = \sup_{0 < \lambda \le Q^{-1}(M_v) \wedge \Lambda} U^+(\lambda). \tag{43}$$

Note that U^+ is continuous on $[0, \Lambda]$ and $\limsup_{\lambda \to 0} U^+(\lambda) \le 0$. Similar to the proof of Lemma **??**, there exists $Q^{-1}(M_v) \wedge \Lambda > \epsilon > 0$ such that

$$\sup_{0 < \lambda \le Q^{-1}(M_v) \wedge \Lambda} U^+(\lambda) = \sup_{\epsilon \le \lambda \le Q^{-1}(M_v) \wedge \Lambda} U^+(\lambda).$$

By Lemma **??**, Lemma **??**, and Assumptions **??** and **??**, U^+ is continuous. Therefore there exists $\lambda_s^* \in [\epsilon, Q^{-1}(M_v) \wedge \Lambda]$ such that $U^+(\lambda_s^*) = \sup_{\epsilon \le \lambda \le Q^{-1}(M_v) \wedge \Lambda} U^+(\lambda)$, and, by (**??**) and (**??**), $U^*(\lambda_s^*) = \sup_{\lambda \ge 0} U^*(\lambda)$.

Using the second derivative of $U^+(\lambda)$, we next show that $U^+(\lambda)$ is concave in λ on $(0, \Lambda]$. First, by (**??**),

$$\frac{dU^+(\lambda)}{d\lambda} = V'(\lambda) - Q(\gamma(\lambda)) - \frac{c}{\gamma(\lambda) - \lambda} - \lambda \left(\frac{c}{(\gamma(\lambda) - \lambda)^2} \right).$$

Also, again by (**??**),

$$\frac{d^2 U^+(\lambda, \gamma(\lambda))}{d\lambda^2} = V''(\lambda) - \frac{2c}{(\gamma(\lambda) - \lambda)^2} - \frac{2c(1 - \gamma'(\lambda))\lambda}{(\gamma(\lambda) - \lambda)^3}.$$

Hence, since $\gamma(\lambda) > \lambda$ by Lemma **??**, and by Lemma **??** and Lemma **??** each term is non-positive,

$$\frac{d^2 U^+(\lambda, \gamma(\lambda))}{d\lambda^2} < 0, \quad \forall \lambda \in (0, \Lambda]. \tag{44}$$

Thus, $U^+(\lambda)$ is strictly concave in λ, for $\lambda \in (0, \Lambda]$. We thus can use the necessary first-order conditions for an optimal solution. Then if $\tilde{\lambda}_s \in [0, \Lambda]$ exists, it must be the optimal solution due to strict concavity. Otherwise, the optimal solution must be at the boundary Λ. The second part of the result follows from (**??**). □

F.3. Proofs of Propositions ?? and ?? and Theorem ??

Proof of Proposition **??***:* Consider two travel burden functions t_1 and t_2 such that $t_2(x) = t_1(x) + a$ for some constant $a \geq 0$, for all $x \geq 0$, and assume that Assumptions **??**, **??**, **??**, and **??** hold and that $R_2^* > 0$. Let V_1 and V_2 be defined as in (**??**) when the travel burden is given by t_1 and t_2 respectively. Therefore, we have $V_i'(\lambda) = m - t_i(F^{-1}(\lambda/\Lambda))$ by (**??**). Let h_i, $i = 1, 2$, be defined by

$$h_i(\lambda) = V_i'(\lambda) + \lambda V_i''(\lambda) - Q(\gamma(\lambda)) - \frac{c}{\gamma(\lambda) - \lambda} - \frac{c\lambda}{(\gamma(\lambda) - \lambda)^2}. \tag{45}$$

By the assumption on t_1 and t_2, Lemma **??**, and (**??**), we have

$$h_1(\lambda) \geq h_2(\lambda). \tag{46}$$

Because $R_2^* > 0$, $\lambda_2^* > 0$ by (**??**) and (**??**). Therefore, by Lemma **??**, $h_2(\lambda_2^*) = 0$, and by (**??**), $h_2(\lambda) > 0$ for $\lambda \in (0, \lambda_2^*)$. Hence, $h_1(\lambda) > 0$ for $\lambda \in (0, \lambda_2^*)$ by (**??**), giving part (i) by Lemma **??**. The proof of $\mu_1^* \geq \mu_2^*$ follows directly from part (i) and Lemmas **??** and **??**.

Now assume that $t_1(x) = 0$. In this case we need to modify the proof of Lemma **??** because the proof relies on Assumption **??**. Our approach can be used to show that we can still find the optimal solution in a similar way. The details are very similar to the case with heterogeneous customers, so we only present the summary of results here.

Consider a service system when $v = m$ with probability 1. In this case the definition of the equilibrium is different because all customers have the same utility from seeking service. In equilibrium customers choose to join the service with probability q such that the utility from seeking service becomes equal to not seeking service. We refer the reader to **?** and **?** for more details.

The rest of the analysis for identifying the optimal actions for a specialist is identical to that in Section **??** by setting $V'(\lambda) = \bar{\Theta}^{-1}(\lambda/\Lambda) = m$ for all $\lambda \in [0, \Lambda]$. Next we provide the details. Let $k(p, \mu) = m - Q(\mu) - \beta p - c/\mu$. For given $\mu > 0$ and $p \geq 0$, define

$$q^*(p, \mu) = \begin{cases} \sup\{q : m - Q(\mu) - \beta p - \frac{c}{\mu - q\Lambda} = 0\}, & \text{if } k(p, \mu) \geq 0 \\ 0, & \text{otherwise} \end{cases}.$$

Then the equilibrium arrival rate $\lambda_m(p, \mu)$ is given by $\lambda_m(p, \mu) = q^*(p, \mu)\Lambda$. The specialist's objective is given by

$$R_m(p, \mu) = \begin{cases} p\lambda_m(p, \mu), & \text{if } p \geq 0 \text{ and } \mu > \lambda(p, \mu) \\ 0, & \text{otherwise} \end{cases}.$$

Let $R_m^* = \sup_{p \geq 0, \mu > 0} R_m(p, \mu)$. As in Section **??**, we can cast the problem in terms of choosing the arrival rate and the service rate. Define

$$p_m(\lambda, \mu) = \sup \left\{ p \in \mathbb{R} : m - Q(\mu) - \beta p - \frac{c}{\mu - \lambda} \leq 0 \right\} \vee 0 \tag{47}$$

and

$$\tilde{R}_m(\lambda,\mu) = \begin{cases} p_m(\lambda,\mu)\lambda, & \text{if } \lambda \in [0,\Lambda] \text{ and } \mu > \lambda \\ 0, & \text{otherwise} \end{cases}.$$

Let $\tilde{R}_m^* = \sup_{\lambda\geq 0,\mu>0} \tilde{R}_m(p,\mu)$. Then Lemma **??** still holds in this case. That is, $\tilde{R}_m^* = R_m^*$.

Lemma **??** and the following result corresponding to Lemma **??** hold. Let $\tilde{\lambda}_m \in [0,\Lambda]$ be a solution to

$$m - Q(\gamma(\tilde{\lambda}_m)) - \frac{c}{\gamma(\tilde{\lambda}_m) - \tilde{\lambda}_m} - \frac{c\tilde{\lambda}_m}{(\gamma(\tilde{\lambda}_m) - \tilde{\lambda}_m)^2} = 0. \tag{48}$$

If there exist $\lambda > 0$ and $\mu > \lambda$ such that $\tilde{R}_m(\lambda,\mu) > 0$, and $\tilde{\lambda}_m \in [0,\Lambda]$ exists, then $\lambda_1^* = \tilde{\lambda}_m$; otherwise $\lambda_1^* = \Lambda$. If no such $\lambda > 0$ and $\mu > \lambda$ exist, then $\tilde{R}_m^* = 0$.

Now assume that t_2 satisfies Assumption **??**, as if not the proof is very similar to the one we present below using (**??**). Because $R_2^* > 0$, $\lambda_2^* > 0$ by (**??**) and (**??**). Also, by (**??**) and (**??**), $p_m(\lambda,\mu) \geq p(\lambda,\mu)$ for any $\lambda > 0$ and $\mu > 0$, which implies that $\tilde{R}_m(\lambda,\mu) \geq \tilde{R}_2(\lambda,\mu)$, where $\tilde{R}_2$ is defined as in (**??**) for the case in which the travel burden is given by t_2. Hence, if $\lambda_2^* > 0$, then $\lambda_1^* > 0$.

Let h_1 be defined by

$$h_1(\lambda) = m - Q(\gamma(\lambda)) - \frac{c}{\gamma(\lambda) - \lambda} - \frac{c\lambda}{(\gamma(\lambda) - \lambda)^2}. \tag{49}$$

Let V_2 be defined as in (**??**) when the travel burden is given by t_2. Also, let h_2 be defined as in (**??**); that is,

$$h_2(\lambda) = V_2'(\lambda) + \lambda V_2''(\lambda) - Q(\gamma(\lambda)) - \frac{c}{\gamma(\lambda) - \lambda} - \frac{c\lambda}{(\gamma(\lambda) - \lambda)^2}. \tag{50}$$

From Assumption **??**, Lemma **??**, and (**??**) and (**??**), $h_1(\lambda) \geq h_2(\lambda)$.

The result then follows from the discussion following (**??**) above. □

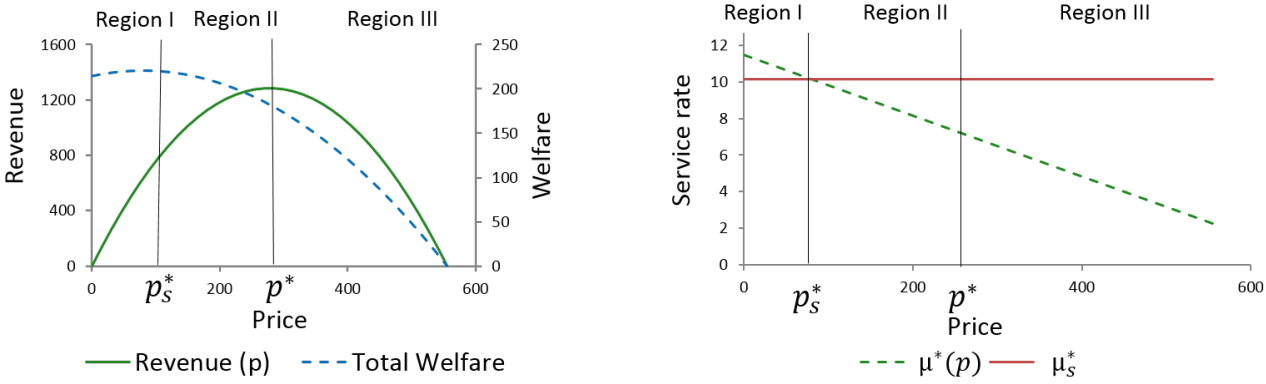
Proof of Theorem **??***:* For parts (i) and (ii), the proof follows from Lemmas **??** and **??**, in a way similar to the proof of Proposition **??**. For the third part of the proposition, if customers are homogeneous in their travel burden (i.e., $t(x)$ is a constant), then the objectives of the welfare-maximizing specialist and the revenue-maximizing specialist are identical by (**??**) and (**??**). Therefore, their optimal actions are identical as well. □

Proof of Proposition **??***:* We solve the optimization problem for the revenue-maximizing specialist and the welfare-maximizing specialist using the approach given in **?**. It can be shown that $p_s^* = \delta(m - 2\sqrt{c\delta})/(\beta(t_0 + 2\delta))$. When the price is exogenous, the optimal arrival rate for a revenue-maximizing specialist is $\lambda^*(p) = (m - p\beta - 2\sqrt{c\delta})/(t_0 + \delta)$, and the optimal service rate is given by

$\mu^*(p) = (m - p\beta)/(t_0 + \delta) + \sqrt{c/\delta}(t_0\delta)(t_0 + \delta)$. The optimal service rate for a welfare-maximizing specialist is given by $\mu_s^* = (m - 2\sqrt{c\delta})/(t_0 + 2\delta) + \sqrt{c/\delta}$. Comparing the two service rates, the price at which the two service rates (and so the arrival rates) are the same is given by $p = p_s^*$. Proposition **??** then follows. □

G. Exogenous Price Setting Numerical Analysis

We better illustrate Proposition **??** using a numerical analysis. We continue to use the linear functional forms introduced above and assume that $\delta = 1, m = 60, t_0 = 5, \beta = 0.1$, and $c = 5$. Figure **????** compares the total revenue and welfare, and Figure **????** the optimal service rate, against a fixed price set by a third party. We can observe that the revenue-maximizing price (p^*) is higher than that of the welfare-maximizing price (p_s^*). We also observe that there is a linear relationship between the fixed price and the optimal service and arrival rates. The threshold mentioned in Proposition **??** is marked by p_s^* in the figures ($p_s^* = 79$), above which the optimal arrival and service rates for a revenue-maximizing specialist are lower than those for a welfare-maximizing specialist. If the exogenous price is in Region I ($< p_s^*$), then not only does the specialist's revenue increase, but the welfare increases as well with a higher price. Similarly, revenue and welfare both increase if the price is reduced in Region III. Hence the third party who determines the exogenous price is limited to choosing a price from Region II (between p_s^* and p^*).



(a) Optimal Revenue and Total Welfare vs. Price: $\delta = 1; m = 60; t_0 = 5; \beta = 0.1; c = 5$

(b) Optimal Service Rates vs. Price: $\delta = 1; m = 60; t_0 = 5; \beta = 0.1; c = 5$

Figure 5 Analysis of optimal service rates, revenue, and total welfare when price per visit is exogenous

H. Proofs of the results in Section ??

H.1. Proof of Proposition ??

Given a strategy for patients (see Section **??** for a mathematical definition), let λ_{in} denote the arrival rate for in-person *mode* patients and λ_{tm} the arrival rate for telemedicine *mode* patients. Let λ_t denote the arrival rate of telemedicine *visits* and λ_i denote the arrival rate of in-person *visits* per unit time, which includes in-person visits from patients opting for the telemedicine mode of treatment as well. Hence,

$$\lambda_i = \lambda_{in} + (1-\alpha)\lambda_{tm} \text{ and } \lambda_t = \alpha\lambda_{tm}. \tag{51}$$

We set $\boldsymbol{\lambda} = (\lambda_i, \lambda_t)$. We next establish the equilibrium for a given p and μ. We start with the following elementary result:

LEMMA 10. *Fix $\lambda, p,$ and $\mu \in \mathbb{R}_+$. If, for a patient located at distance x_1, $\Psi_i(\lambda,p,\mu,x_1) \geq \Psi_t(\lambda,p,\mu,x_1)$, then $\Psi_i(\lambda,p,\mu,x) > \Psi_t(\lambda,p,\mu,x)$ for all $x < x_1$. Similarly, if, for a patient located at distance x_1, $\Psi_i(\lambda,p,\mu,x_1) \leq \Psi_t(\lambda,p,\mu,x_1)$, then $\Psi_i(\lambda,p,\mu,x) < \Psi_t(\lambda,p,\mu,x)$ for all $x > x_1$. Also, if $\Psi_t(\lambda,p,\mu,x_1) \leq 0$, then $\Psi_t(\lambda,p,\mu,x) < 0$ for all $x > x_1$.*

The proof immediately follows from the definitions of Ψ_t and Ψ_i in (**??**) and (**??**) and the fact that t is strictly increasing. We can then deduce that for a strategy to be an equilibrium it has to have the following structure: There exist $x_1 \geq x_0 \geq 0$ such that patients located closer than x_0 prefer the in-person mode, patients located between x_0 and x_1 prefer the telemedicine mode, and patients located beyond x_1 do not seek treatment.

From here on we assume that $x_{ID} > 0$ (see (**??**)); that is, at least some patients prefer the in-person mode. Otherwise, none of the patients prefers the in-person mode, and this case reduces to a specialist offering only the telemedicine mode, which can be analyzed as in Section **??**. Let $D : \mathbb{R}_+^2 \to \mathbb{R}_+$ be defined as follows:

$$D(p,\mu) = \begin{cases} \inf\{x \geq 0 : k_1(p,\mu,\Lambda F(x)) - t(x) \leq 0 \text{ and } \mu \geq \Lambda F(x)\}, & \text{if } k_1(p,\mu,0) \geq 0 \\ 0, & \text{otherwise} \end{cases}, \tag{52}$$

where $k_1(p,\mu,\ell) = m_i - Q(\mu) - \beta p - c/(\mu - \ell)$. Thus, $D(p,\mu)$ denotes the location of the patient who is indifferent between choosing treatment and not seeking treatment when the telemedicine mode is not available. Also, given $p \geq 0$ and $\mu \geq 0$, let the telemedicine mode utility component independent of patient location be given by $k_2(p,\mu,\ell) = \alpha m_t + (1-\alpha)m_i - s - Q(\mu) - \beta p - c/(\mu - \ell)$, and define $x_{TM} : \mathbb{R}_+^2 \to \mathbb{R}_+$ by

$$x_{TM}(p,\mu) = \begin{cases} \inf\{x \geq 0 : k_2(p,\mu,\Lambda F(x)) - (1-\alpha)t(x) \leq 0 \text{ and } \mu > \Lambda F(x)\}, & \text{if } k_2(p,\mu,0) \geq 0 \\ 0, & \text{otherwise} \end{cases}. \tag{53}$$

Therefore, $x_{TM}(p,\mu)$ denotes the location of the patient who is indifferent between choosing the telemedicine mode and not seeking treatment when the telemedicine option is available. Let i, t, and o denote three alternatives: the in-person mode, the telemedicine mode, and not seeking treatment, respectively. Hence the action set for the patients is given by $\mathcal{A} = \{i, t, o\}$. Let $g_x(a)$ denote the probability that a patient in location $x \in \mathbb{R}^+$ chooses action $a \in \mathcal{A}$. We have the following result, which implies Proposition **??**. We need the more general result in proving Theorems **??** and **??** below.

THEOREM 4. *For fixed $p \geq 0$ and $\mu > 0$,*

i) if $D(p,\mu) \leq x_{ID}$, then the unique[7] *equilibrium is given by $g_x(i) = 1$ for $0 \leq x \leq D(p,\mu)$ and $g_x(o) = 1$ for $x > D(p,\mu)$;*

ii) if $D(p,\mu) > x_{ID}$, then $x_{ID} < x_{TM}(p,\mu)$, and the unique equilibrium is given by $g_x(i) = 1$ for $0 \leq x \leq x_{ID}$, $g_x(t) = 1$ for $x \in (x_{ID}, x_{TM}(p,\mu)]$, and $g_x(o) = 1$ for $x > x_{TM}(p,\mu)$.

Based on Theorem **??**, for a given (p,μ), if the condition in part (i) holds ($D(p,\mu) \leq x_{ID}$), then patients do not prefer the telemedicine mode and none of the patients seeks treatment if in addition $D(p,\mu) = 0$; if the condition in part (ii) holds ($D(p,\mu) > x_{ID}$), some of the patients choose the in-person mode and some choose the telemedicine mode.

Proof of Theorem **??***:* First, by Lemma **??** and (**??**),

$$\Psi_i(\lambda, p, \mu, x) > \Psi_t(\lambda, p, \mu, x), \text{ for all } x < x_{ID}, \tag{54}$$

for any λ, p, and μ. Hence in any equilibrium

$$g_x(t) = 0 \text{ for } 0 \leq x < x_{ID}. \tag{55}$$

Fix $p \geq 0$ and $\mu > 0$ for the rest of the proof.

Part (i): Assume that $x_{TM} \leq D(p,\mu) \leq x_{ID}$. Let $G = \{g_x : x \geq 0\}$ denote the strategy profile given in part (i) and λ_G denote the associated total arrival rate defined as in (**??**). Note that $\lambda_G = \Lambda F(D(p,\mu))$ by (**??**). We first show that G is an equilibrium. By Lemma **??**, (**??**), and (**??**), for $x \in [0, D(p,\mu)]$, (**??**) holds with $g_x(i) = 1$. By (**??**), (**??**), and (**??**), (**??**) holds with $g_x(o) = 1$ for $x \in (D(p,\mu), x_{ID}]$. Finally, by Lemma **??** and (**??**), (**??**) holds with $g_x(o) = 1$ for $x > x_{ID}$. Hence G is an equilibrium.

Next we show $G = \{g_x : x \geq 0\}$ is the unique equilibrium. Let $G' = \{g'_x : x \geq 0\}$ be another equilibrium, and denote the associated total arrival rate by λ', defined as in (**??**). First assume that

[7] in almost everywhere (a.e.) measure theory technical sense.

$g'_{x'}(t) > 0$ for some x', so $\Psi_t(\lambda, p, \mu, x') \geq 0$ by (**??**). Then, by (**??**), $x' \geq x_{ID}$. Also, by $\Psi_t(\lambda, p, \mu, x') \geq 0$, (**??**), and the fact that t is strictly increasing,

$$\Psi_t(\lambda', p, \mu, x) > 0 \text{ for all } x < x', \tag{56}$$

and by (**??**) this implies that

$$\Psi_i(\lambda', p, \mu, x) > 0 \text{ for all } x < x_{ID}. \tag{57}$$

If $x' = x_{ID}$, then G and G' must be equal a.e. by (**??**) and (**??**). Thus assume that $x' > x_{ID}$. Then, again by (**??**) and (**??**), $g'_x(i) = 1$ for all $x < x_{ID}$, and $g'_x(t) = 1$ for $x \in (x_{ID}, x')$. Therefore $\lambda' > \lambda$ by (**??**) and the fact that $x' > x_{ID}$. So by (**??**) there exists $x'' < D(p, \mu)$ such that

$$\Psi_i(\lambda', p, \mu, x) < 0 \text{ for all } x \in [x'', D(p, \mu)]. \tag{58}$$

Since $D(p, \mu) \leq x_{ID}$ and $x' > x_{ID}$, (**??**) contradicts (**??**); therefore no such x' exists and patients only seek in-person visits. It then readily follows from Lemma **??** that G is the unique equilibrium (in the a.e. sense).

Part (ii): Now assume that $D(p, \mu) > x_{ID}$. First we prove that $x_{TM}(p, \mu) > x_{ID}$. Note that for $x \in (x_{ID}, D(p, \mu))$,

$$\Psi_t(\Lambda F(D(p, \mu)), p, \mu, x) \geq \Psi_i(\Lambda F(D(p, \mu)), p, \mu, x) > 0, \tag{59}$$

where the first inequality follows from the definition of x_{ID} and the second follows from the definition of $D(p, \mu)$. By (**??**), (**??**) implies that $x_{TM}(p, \mu) \geq D(p, \mu) > x_{ID}$.

Let $G = \{g_x : x \geq 0\}$ denote the strategy profile given in part (ii) and λ_G denote the associated total arrival rate defined as in (**??**). We first show that G is an equilibrium. Note that $\lambda_G = \Lambda F(x_{TM}(p, \mu))$. By the definition of x_{ID} and the fact that $x_{TM}(p, \mu) > x_{ID}$, $\Psi_i(\lambda_G, p, \mu, x) \geq \Psi_t(\lambda_G, p, \mu, x) > 0$ for all $x \in [0, x_{ID}]$. Hence, for $x \in [0, x_{ID}]$, (**??**) holds with $g_x(i) = 1$. Similarly, for $x \in [x_{ID}, x_{TM}(p, \mu)]$, by (**??**) and (**??**), $\Psi_t(\lambda_G, p, \mu, x) \geq \Psi_i(\lambda_G, p, \mu, x)$ and $\Psi_t(\lambda_G, p, \mu, x) \geq 0$. Hence, for $x \in [x_{ID}, x_{TM}(p, \mu)]$, (**??**) holds with $g_x(t) = 1$. Finally, by (**??**), the fact that $x_{TM}(p, \mu) > x_{ID}$, and Lemma **??**, $\Psi_i(\lambda_G, p, \mu, x) \leq \Psi_t(\lambda_G, p, \mu, x) < 0$ for all $x > x_{TM}(p, \mu)$. Thus, for $x > x_{TM}(p, \mu)$, (**??**) holds with $g_x(o) = 1$.

Next we show that $G = \{g_x : x \geq 0\}$ is the unique equilibrium. Let $G' = \{g'_x : x \geq 0\}$ be another equilibrium, and denote the associated total arrival rate by λ' defined as in (**??**). First we show that for $y > x_{TM}(p, \mu)$, in any equilibrium $g'_y(o) = 1$. If not, then by (**??**) $\Psi_t(\lambda', p, \mu, y) \geq 0$, and so by Lemma **??**

$$\Psi_t(\lambda', p, \mu, x) > 0, \text{ for all } x < y. \tag{60}$$

This implies that $g'_x(t) = 1$ for all $x \in (x_{ID}, y)$, so $\lambda' > \lambda_G$. But then, by (**??**) and the fact that $x_{TM}(p,\mu) > x_{ID}$, $\Psi_i(\lambda', p, \mu, x) < \Psi_t(\lambda', p, \mu, x) < 0$, for $x \in (x_{TM}, y)$. This obviously contradicts (**??**), so no such y exists. This also implies that $\lambda \leq \lambda_G$. Therefore, by (**??**), (**??**), and Lemma **??**, any equilibrium strategy g'_x must satisfy $g'_x(i) = 1$ for all $x \in [0, x_{ID})$, since $\Psi_i(\lambda', p, \mu, x) > \Psi_t(\lambda', p, \mu, x) > 0$, for all $x \in [0, x_{ID})$, where the last inequality follows from the fact that $x_{ID} < x_{TM}(p,\mu)$. Then it readily follows that $g'_x(t) = 1$ for $x \in (x_{ID}, x_{TM}(p,\mu))$, proving uniqueness a.e. □

H.2. Proof of Theorem ??

We first show that when the specialist does not differentiate between in-person and telemedicine patients, his objective function is separable in a certain sense. This allows us to obtain a closed-form solution for the specialist's optimal decisions. The separability results from the fact that the difference between the utilities (see (**??**)) only depends on the patient's characteristics, not the specialist's decisions.

Before we state the main result in this section we note that if at equilibrium all patients seeking treatment choose the in-person mode or all patients seeking treatment choose the telemedicine mode, then the specialist ends up offering only one mode of treatment: the in-person mode or the telemedicine mode, respectively. His (revenue-maximizing) optimal actions are determined using Lemma **??** and Lemma **??** in Appendix **??**. With a slight abuse of notation, we let (p_k^*, μ_k^*) denote the optimal decisions for a specialist offering only the in-person mode for $k = 1$ and only the telemedicine mode for $k = 2$ for notational simplicity. (Quantities p_1^* and μ_1^* are the optimal choices of the revenue-maximizing provider we identified in Section **??**.) We similarly use subscript k with the function λ defined in (**??**) and R defined as in (**??**) to denote the associated quantities with these two types of specialists. By (**??**), given $\mu \in \mathbb{R}_+$ and $p \in \mathbb{R}_+$,

$$\lambda_1(p,\mu) = \Lambda F(D(p,\mu)) \text{ and } \lambda_2(p,\mu) = \Lambda F(x_{TM}(p,\mu)). \tag{61}$$

We set $\lambda_k^* = \lambda_k(p_k^*, \mu_k^*)$ and $R_k^* = R_k(p_k^*, \mu_k^*)$, $k = 1, 2$, for notational brevity.

PROPOSITION 4. *For a specialist offering both treatment modes, $\bar{R}^* = \max\{R_1^*, R_2^*\}$. For $m = \arg\max_{k=1,2}\{R_k^*\}$ (which denotes the mode whose revenue is higher), $\tilde{p} = p_m^*$ and $\tilde{\mu} = \mu_m^*$. Hence the total arrival rate satisfies $\lambda_i(\tilde{p}, \tilde{\mu}) + \lambda_t(\tilde{p}, \tilde{\mu}) = \lambda_m^*$.*

From Proposition **??**, in order to find the optimal decisions for a specialist offering both modes of treatment, we only need to find the optimal actions for a specialist offering each mode exclusively (to a patient population with the same characteristics). The specialist offering both modes of treatment

will choose the same action as the one that obtains more revenue. The resultant equilibrium may or may not be a mixed equilibrium, as we explain next.

We can also identify the ensuing equilibrium if the specialist chooses the optimal actions given in Proposition **??**. Specifically, if $R_1^* \geq R_2^*$ and the specialist chooses the parameters $\tilde{p} = p_1^*$ and $\tilde{\mu} = \mu_1^*$, then patients do not choose the telemedicine mode in the equilibrium.

If $R_1^* < R_2^*$, then $D(p_2^*, \mu_2^*) > x_{ID}$, and so the characterization of the equilibrium follows from Theorem **??**(ii).

Proof of Proposition **??***:* We prove the result by showing that for any $p \geq 0$ and $\mu > 0$,

$$\bar{R}(p,\mu) = \max\{R_1(p,\mu), R_2(p,\mu)\}. \tag{62}$$

Fix $p \geq 0$ and $\mu > 0$. Note that if $R_i(p,\mu) > (\geq) R_{i'}(p,\mu)$, then $\lambda_i(p,\mu) > (resp., \geq)\lambda_{i'}(p,\mu)$ by (**??**), for $i \in \{1,2\}$ and $i' = \{1,2\} \setminus \{i\}$. We prove the following result below.

LEMMA 11. *If $\lambda_1(p,\mu) \geq \lambda_2(p,\mu)$, then $D(p,\mu) \leq x_{ID}$, and if $\lambda_1(p,\mu) < \lambda_2(p,\mu)$, then $D(p,\mu) > x_{ID}$.*

Hence if $R_1(p,\mu) \geq R_2(p,\mu)$, then by Theorem **??**(i), (**??**), and Lemma **??**, $\bar{R}(p,\mu) = R_1(p,\mu)$. If on the other hand $R_1(p,\mu) < R_2(p,\mu)$, then by Theorem **??**(ii), (**??**), and Lemma **??**, $\bar{R}(p,\mu) = R_2(p,\mu)$, proving (**??**). We complete the proof by proving Lemma **??**.

Assume that

$$\lambda_1(p,\mu) \geq \lambda_2(p,\mu). \tag{63}$$

This implies by (**??**) that

$$D(p,\mu) \geq x_{TM}(p,\mu). \tag{64}$$

Therefore

$$\Psi_i(\lambda_1(p,\mu), p, \mu, D(p,\mu)) \geq \Psi_t(\lambda_2(p,\mu), p, \mu, D(p,\mu)) \geq \Psi_t(\lambda_1(p,\mu), p, \mu, D(p,\mu)), \tag{65}$$

where the first inequality follows from (**??**), (**??**), and (**??**), and the second inequality follows from (**??**). By Lemma **??** and (**??**), $\Psi_i(\lambda_1(p,\mu), p, \mu, x) \geq \Psi_t(\lambda_1(p,\mu), p, \mu, x)$, for all $x \leq D(p,\mu)$. Thus $D(p,\mu) \leq x_{ID}$ by (**??**), proving the first part of the lemma.

Now assume that

$$\lambda_1(p,\mu) < \lambda_2(p,\mu). \tag{66}$$

This implies by (**??**) that $D(p,\mu) < x_{TM}(p,\mu)$. Therefore

$$\Psi_i(\lambda_1(p,\mu),p,\mu,D(p,\mu)) < \Psi_t(\lambda_2(p,\mu),p,\mu,D(p,\mu)) \leq \Psi_t(\lambda_1(p,\mu),p,\mu,D(p,\mu)), \tag{67}$$

where the first inequality follows from (**??**), (**??**), (**??**), and Lemma **??**, and the second inequality follows from (**??**). By Lemma **??** and (**??**), $\Psi_i(\lambda_1(p,\mu),p,\mu,x) < \Psi_t(\lambda_1(p,\mu),p,\mu,x)$ for all $x \geq D(p,\mu)$. Thus $D(p,\mu) > x_{ID}$ by (**??**), proving the second part of the lemma and concluding the proof. □

The proof of Theorem **??** follows from Theorem **??** and Proposition **??**. Assume that $\lambda^* > 0$ and (**??**) holds. By (**??**), $\lambda_1^* = \Lambda F\left(D(p_1^*,\mu_1^*)\right)$. Then, by (**??**), (**??**), and (**??**), $\Psi_t(\lambda_1^*,p_1^*,\mu_1^*,D(p_1^*,\mu_1^*)) > \Psi_i(\lambda_1^*,p_1^*,\mu_1^*,D(p_1^*,\mu_1^*)) = 0$. Thus $x_{TM}(p_1^*,\mu_1^*) > D(p_1^*,\mu_1^*)$. Therefore, by (**??**), $R_2^* > R_1^*$. Then, by Proposition **??**, $\bar{R}^* = R_2^*$ and $\tilde{\lambda} = \lambda_2^*$, proving part (i). Part (ii) follows from Lemma **??**. □

H.3. Proof of Theorem ??

Note that it is enough to prove that

$$U_d(\lambda_2^*,\mu_2^*) \geq U(\lambda_1^*,\mu_1^*), \tag{68}$$

where λ_i^* and μ_i^* are defined as in the previous section for $i = 1,2$. This follows from the fact that when telemedicine is feasible the optimal arrival and service rates coincide with those when the specialist offers only the telemedicine mode by Theorem **??**.

If $\lambda_1^* = 0$, then $U(\lambda_1^*,\mu_1^*) = 0$ by (**??**) as well, so (**??**) holds trivially. Assume that $\lambda_1^* > 0$ and that (**??**) holds. If $U_d(\lambda,\mu) = 0$ for all $\lambda > 0$ and $\mu > 0$, then, by (**??**) and (**??**), $U_d(\lambda,\mu) = 0$ for all $\lambda > 0$ and $\mu > 0$ as well (recall that we assume $x_{ID} > 0$). So also assume that $U_d(\lambda,\mu) > 0$ for some $\lambda > 0$ and $\mu > 0$.

Let $U_d^*(\lambda) = \sup_{\mu>\lambda} U_d(\lambda,\mu)$. Similar to Lemma **??**, it can be shown that if there exists $\mu > \lambda$ such that $U_d(\lambda,\mu) > 0$, then $U_d^*(\lambda) = U_d(\lambda,\gamma(\lambda))$.

If the specialist chooses to serve arrival rate λ, then the distance of the farthest patient is given by $x = F^{-1}(\lambda/\Lambda)$. By the proof of Theorem **??**, if (**??**) holds, then $\lambda_1(\mu_1^*,p_1^*) \geq \Lambda F(x_{ID})$. Therefore, by Proposition **??**, Lemma **??**, and (**??**),

$$U_1^*(\Lambda F\left(x_{ID}\right)) \geq U_1^*(\Lambda F\left(x\right)) \text{ for } x \leq x_{ID}. \tag{69}$$

Because $U_d^*(\Lambda F\left(x\right)) = U_1^*(\Lambda F\left(x\right))$ for $x \leq x_{ID}$, we have by (**??**) that

$$U_d^*(\Lambda F\left(x_{ID}\right)) \geq U_1^*(\Lambda F\left(x\right)) \text{ for } x \leq x_{ID}. \tag{70}$$

Also, by (**??**), $dV_d^*(\lambda)/d\lambda = \tilde{m} - (1-\alpha)t(F^{-1}(\lambda/\Lambda))$, for $\lambda > \Lambda F(x_{ID})$. Therefore, by (**??**),

$$\frac{dU_d^*(\lambda)}{d\lambda} = \frac{dU_2^*(\lambda)}{d\lambda}, \text{ for } \lambda > \Lambda F(x_{ID}). \tag{71}$$

By Proposition **??**, Lemma **??**, and (**??**), $dU_2^*(\lambda)/d\lambda > 0$ for $\lambda < \lambda_2^*$. Thus, by (**??**), U_d^* is increasing for $\lambda \in (\Lambda F(x_{ID}), \lambda_2^*)$. Combined with (**??**), this gives the desired result. □

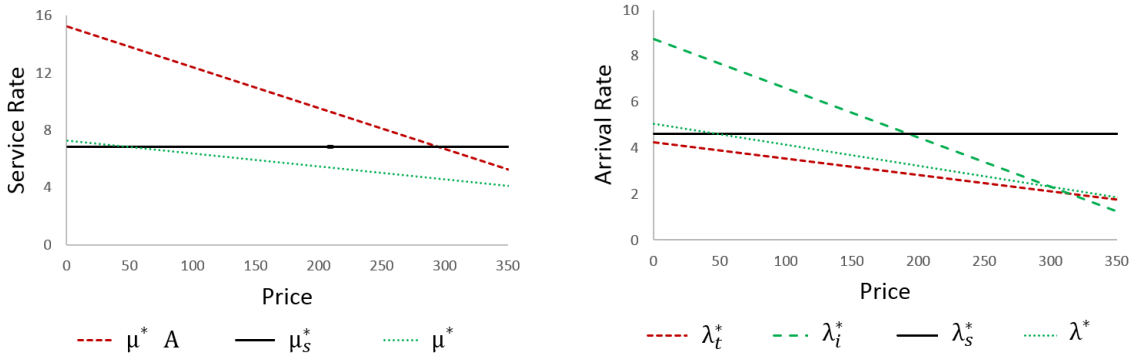
H.4. Telemedicine equilibrium when prices are exogenous

In this section, we analyze the impact of introducing telemedicine when the prices are fixed exogenously (recall that we analyze the same case without telemedicine in Section **??**). We assume a setting similar to Section **??**; the specialist employs the same service rate for both in-person and telemedicine visits. The model parameters specific to telemedicine are explained in Section **??**. We also assume the exogenous price to be the same for both types of visits. We make use of the simplifying assumptions on page **??** in an attempt to obtain closed-form solutions to get better insights. Despite the simplifying assumptions and only one decision variable for the specialist, namely the service rate, finding a closed-form solution for the equilibrium arrival rates is analytically cumbersome. Hence we optimize the specialist's revenue assuming specific values for the model parameters and then obtain insights. We assume the following values for the model parameters: $\delta = 1, m_i = 60, m_t = 60, t_0 = 10, \beta = 0.1, c = 5, \alpha = 0.75$, and $s = 10$. We then find the equilibrium arrival rates as functions of the service rate, μ. Then we optimize the revenue function to find the optimal service rate. Figures **????** and **????** show the optimal service rates for the specialist and the optimal arrival rates at equilibrium as functions of the exogenous price respectively.

We observe that, similar to Theorem **??**, the total arrival rate and service rate increase with the introduction of telemedicine. Similar to Proposition **??**, there is a threshold for the exogenous price beyond which the service rates after the introduction of telemedicine are lower than those of the welfare-maximizing specialist. If the exogenous price is lower, it will tend to drive the optimal service rates higher. Also, this threshold has moved to the right after telemedicine; that is, the threshold is higher than that when telemedicine mode is not available, so if the existing price continues after telemedicine, service rates will be higher and may be higher than those of the welfare-maximizing specialist.

I. Additional Numerical Analysis

In this section we extend our numerical analysis in Section **??**. We first look at two important measures, utilization and congestion costs, and see how sensitive they are to the clinical feasibility of telemedicine (α) and the transportation cost (t_0). We then look at the sensitivity of optimal



(a) Optimal Service Rates vs. Price: $\delta = 1; m_i = 60; m_t = 60; t_0 = 10; \beta = 0.1; c = 5; \alpha = 0.75; s = 10$

(b) Optimal Arrival Rates vs. Price: $\delta = 1; m_i = 60; m_t = 60; t_0 = 10; \beta = 0.1; c = 5; \alpha = 0.75; s = 10$

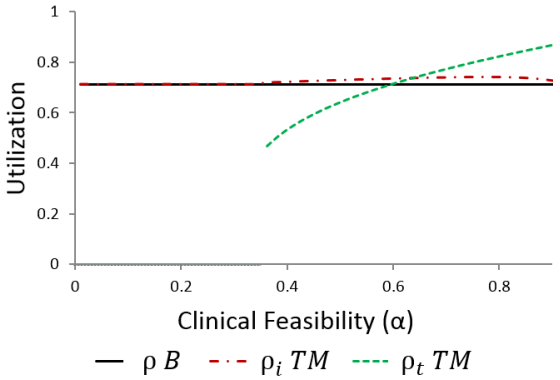
Figure 6 **Analysis of post-telemedicine optimal service rate and arrival rates when price per visit is exogenous: $\mu^* A$—optimal service rates for a revenue-maximizing specialist after telemedicine; λ_i^* and λ_t^*—optimal arrival rates for in-person visits and telemedicine visits after telemedicine; μ_s^* and λ_s^*—optimal service and arrival rates for a welfare-maximizing specialist before telemedicine; μ^* and λ^*—optimal service and arrival rates for a revenue-maximizing specialist before telemedicine**

service rates, arrival rates, revenue, optimal prices, total welfare, and total utility with respect to the relative benefit of telemedicine visits (m_t/m_i) and the setup cost (s).

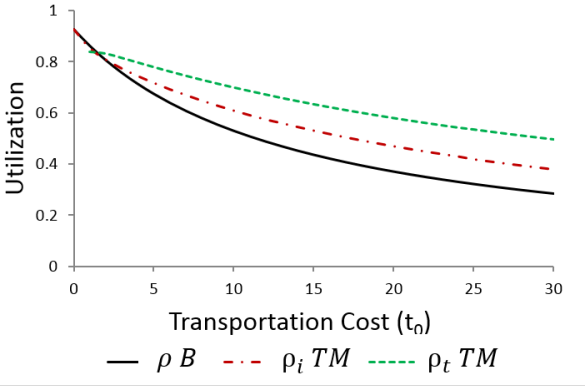
Figures **????** and **????** compare the specialist's utilization before telemedicine (ρ B) and after telemedicine (ρ_i TM and ρ_t TM). Utilization goes up with the clinical feasibility of telemedicine, with more patients getting treatment than before. Hence the congestion cost for all patients increases when the specialist can choose two different service rates for in-person and telemedicine visits. Similar to what we observed in Figure **????**, utilization for in-person visits is higher than utilization for telemedicine visits if α is not too high ($0.3 \leq \alpha \leq 0.6$). This is because the service rates for telemedicine visits are higher (Figure **????**), and with α being not so high demand (λ) is lower as well. However, when α is higher ($\alpha > 0.6$), utilization for telemedicine visits is higher than that of in-person visits due to higher demand.

Even though the optimal service rate decreases with the travel burden (Figure **????**), it is interesting to note that the utilization also drops with the travel burden (Figure **????**). This is because the increased travel burden decreases patient utility and hence their arrival rate.

Figures **????** and **????** compare the congestion costs at optimal values before telemedicine for a revenue-maximizing specialist (Cong B) and a welfare-maximizing specialist (Cong S) and after telemedicine for in-person visits (Cong i) and for telemedicine visits (Cong TM), and the weighted congestion cost (α Cong TM + $(1-\alpha)$ Cong i), as functions of α in Figure **????** and t_0 in Figure **????**. There is no difference between the congestion cost before telemedicine (Cong B) and the



(a) Utilization vs. Clinical Feasibility of Telemedicine (α): $m_t = 60; t_0 = 4$

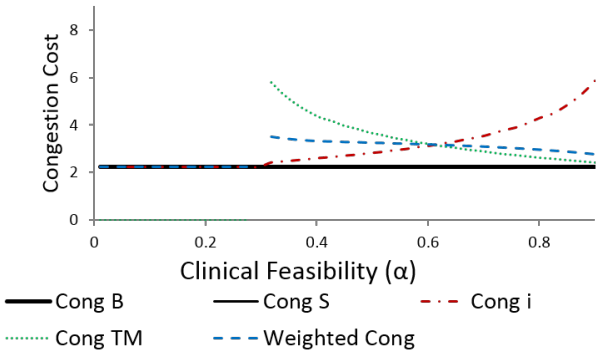


(b) Utilization vs. Transportation Cost ($\boldsymbol{t_0}$): $\alpha = 0.75; m_t = 60$

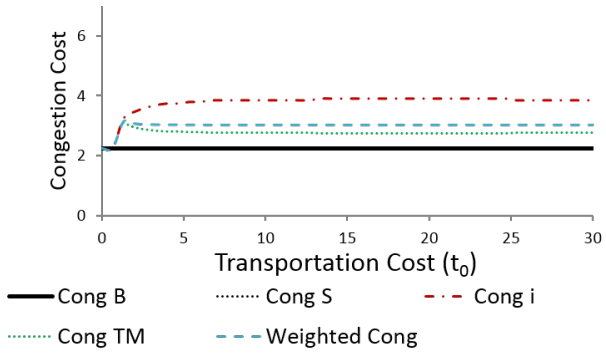
Figure 7 **Analyzing the utilization of the specialist at optimal values, optimal prices, and optimal revenue:** $\delta = 1; m_i = 60; \beta = 0.1; s = 10; c = 5$

congestion cost for a welfare-maximizing specialist (Cong S). Even though the optimal arrival rates and service rates are both different in these two cases, their differences ($\mu - \lambda$) are the same and hence the congestion costs are also the same. These lines therefore overlap in Figures **????** and **????**.

We can also observe that congestion costs increase after the introduction of telemedicine owing to greater patient coverage and higher specialist utilization. Because telemedicine also increases productivity by increasing the service rate, as telemedicine becomes more feasible (higher α), congestion costs decrease and approach the value before the introduction of telemedicine.

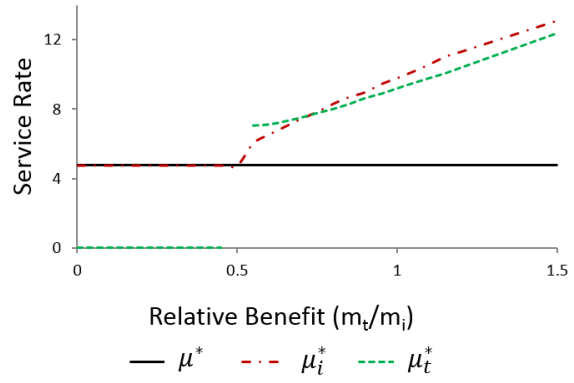


(a) Congestion Costs vs. Clinical Feasibility of Telemedicine ($\boldsymbol{\alpha}$): $m_t = 60; t_0 = 10$

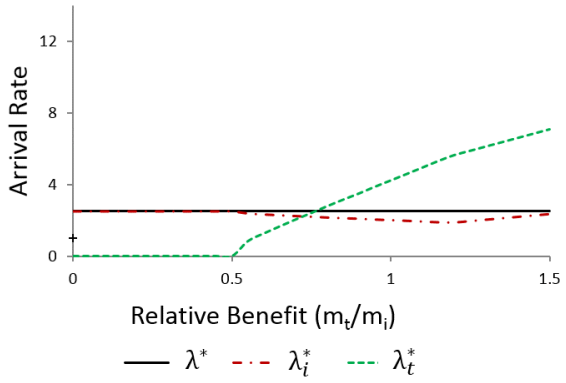


(b) Congestion Costs vs. Transportation Cost ($\boldsymbol{t_0}$): $m_t = 60; \alpha = 0.75$

Figure 8 **Analyzing the total cost of congestion:** $\delta = 1; m_i = 60; \beta = 0.1; s = 10; c = 5$

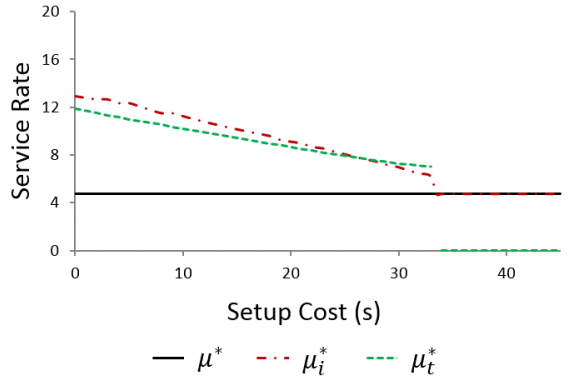


(a) Service Rates vs. Relative Benefit of Telemedicine Visits ($\boldsymbol{m_t/m_i}$): $\alpha = 0.75; t_0 = 10$

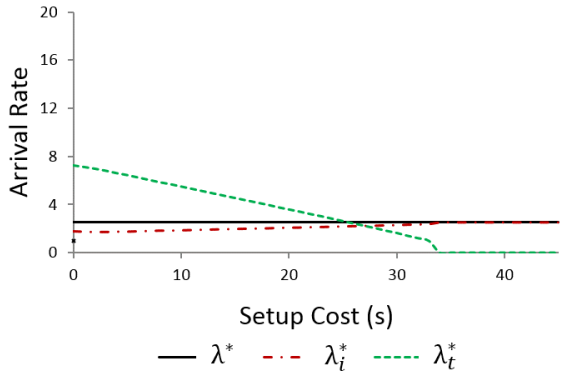


(b) Arrival Rates vs. Relative Benefit of Telemedicine Visits ($\boldsymbol{m_t/m_i}$): $\alpha = 0.75; t_0 = 10$

Figure 9 **Analyzing the optimal service rates (μ^*, μ_i^*, and μ_t^*) and the optimal arrival rates (λ^*, λ_i^*, and λ_t^*):** $\delta = 1; m_i = 60; \beta = 0.1; s = 10; c = 5$



(a) Service Rates vs. Setup Cost of Telemedicine ($\boldsymbol{s}$): $\alpha = 0.8; t_0 = 10$



(b) Arrival Rates vs. Setup Cost of Telemedicine ($\boldsymbol{s}$): $\alpha = 0.8; t_0 = 10$

Figure 10 **Analyzing the optimal service rates (μ^*, μ_i^*, and μ_t^*) and the optimal arrival rates (λ^*, λ_i^*, and λ_t^*):** $\delta = 1; m_i = 60; \beta = 0.1; s = 10; c = 5$

Figures **????** and **????** show how the optimal service rates and arrival rates change with the relative benefit of telemedicine visits (m_t/m_i). We observe that as the relative benefit increases, the provider's efficiency increases as well. When m_t is low, telemedicine is not operationally and economically viable.

Figures **????** and **????** show how the optimal service rates and arrival rates change with the setup cost of telemedicine (s). We observe that as the setup cost increases, both the rates decrease as well. When s is high, telemedicine is not operationally and economically viable.

Figures **????** and **????** compare the optimal prices for the specialist before telemedicine (p^*)

and after telemedicine (p_i^* for in-person visits, p_t^* for telemedicine visits; the weighted price is $\alpha p_t^* + (1-\alpha)p_i^*$) as functions of the relative benefit of telemedicine visits, m_t/m_i (Figure **????**), and the clinical feasibility of telemedicine, α (Figure **????**), respectively. The vertical lines in the following figures are used to separate the different regions. Region I is when the specialist does not offer the telemedicine mode, and the other regions are described as and when needed.

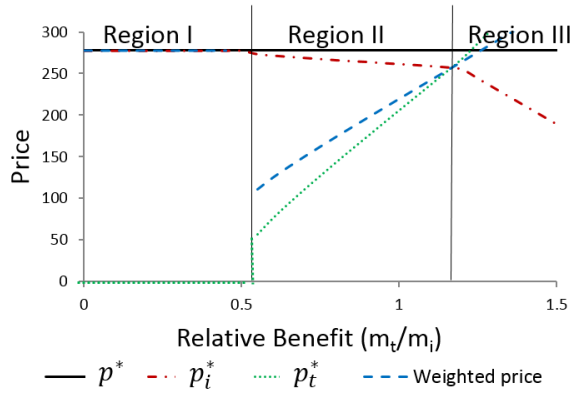
We see that both in-person and telemedicine prices in general drop after the introduction of telemedicine, as seen in Region II. The weighted optimal price for telemedicine patients therefore drops as soon as telemedicine becomes feasible ($m_t/m_i \geq 0.55$) and then increases with telemedicine's value. This is especially true when the benefits are equivalent ($m_t/m_i = 1$). Only if the benefit from telemedicine visits is significantly larger than the benefit from in-person visits ($m_t/m_i \geq 1.2$, Region III, Figure **????**) is it optimal for the specialist to charge higher prices for telemedicine visits. Thus, in Region III, the specialist seems to capture more of the patient surplus by charging a higher price.

The specialist plays with the two levers, price and service rate, to maximize his revenue. For low values of α, the price for telemedicine visits is relatively lower (see Figure **????**). Hence the service rates are higher to optimize revenue. For higher values of α, prices are higher and compensated by a decrease in the service rate.

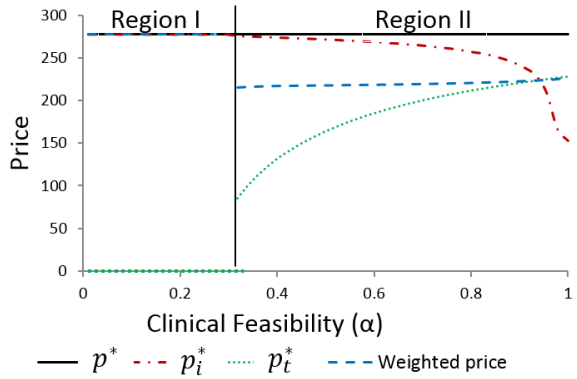
The increased productivity and the increased volume will drive the price difference between telemedicine and in-person visits. The actual prices will of course be influenced by the parameter values. Region II in Figure **????** is also the area where the specialist begins to find the introduction of telemedicine beneficial. When m_t is lower (Region I), telemedicine is not operationally and economically viable.

Figures **????** and **????** compare the optimal revenue (left vertical axis) for the specialist before telemedicine (R^*) and after telemedicine (R^* TM), and also show the proportion of time spent on in-person visits, r (right vertical axis), as functions of the relative benefit of telemedicine visits, m_t/m_i (Figure **????**), and the setup cost of telemedicine, s (Figure **????**). m_t/m_i is allowed to vary from 0 to 1.5. We observe that as the relative benefit of telemedicine visits (m_t/m_i) increases, the proportion of time spent on in-person visits (r) decreases. In other words, as telemedicine becomes more attractive to the patient, the specialist also finds it beneficial to adopt telemedicine.

R^*, the specialist's revenue before telemedicine, is constant with respect to m_t/m_i. In Figure **????** we see that, even when the clinical efficacy of telemedicine is somewhat smaller ($m_t/m_i < 1$), patients would still find telemedicine desirable because of the reduced travel burden and hence the specialist would be introducing the telemedicine mode. This is one of the reasons why telemedicine

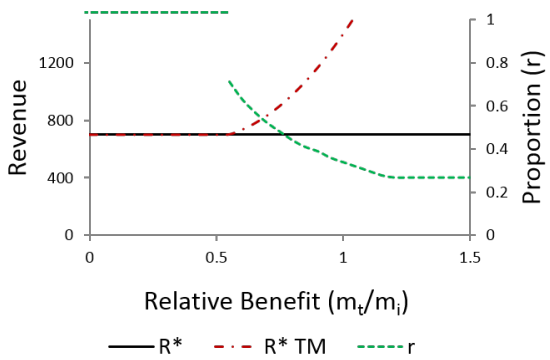


(a) Optimal Prices vs. Relative Benefit of Telemedicine Visits ($\boldsymbol{m_t/m_i}$): $\alpha = 0.75; t_0 = 10$

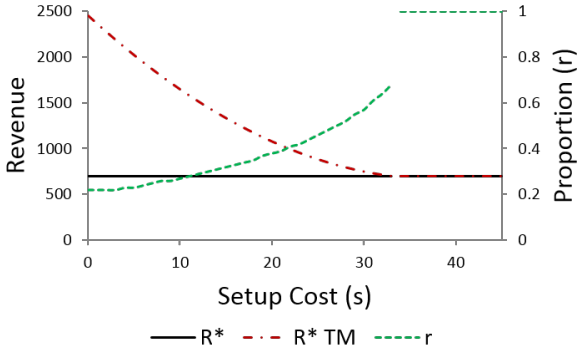


(b) Optimal Prices (p^*, p_i^* and p_t^*) vs. Clinical Feasibility of Telemedicine (α): $\delta = 1; m_i = 60; \beta = 0.1; s = 10; c = 5; m_t = 60; t_0 = 10; s = 10$

Figure 11 **Analyzing the optimal prices (p^*, p_i^*, and p_t^*):** $\delta = 1; m_i = 60; \beta = 0.1; s = 10; c = 5$



(a) Optimal Revenues (R^* and R^*TM) and Proportion of Time Spent on In-person Visits ($\boldsymbol{r}$) vs. Relative Benefit of Telemedicine Visits ($\boldsymbol{m_t/m_i}$): $\alpha = 0.75; t_0 = 10; s = 10$



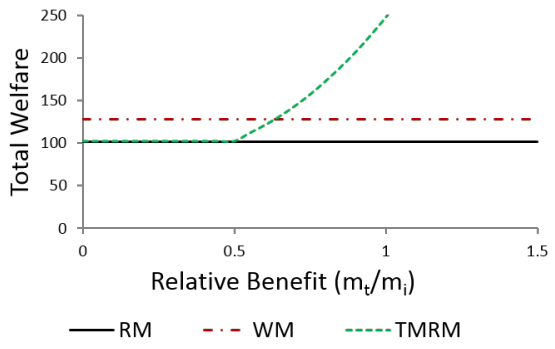
(b) Optimal Revenues (R^* and R^*TM) vs. Setup Cost of Telemedicine ($\boldsymbol{s}$): $\alpha = 0.8; t_0 = 10; m_t = 60$

Figure 12 **Analyzing the optimal revenue of the specialist at optimal values (R^* and R^*TM) and proportion of time spent on in-person visits ($\boldsymbol{r}$):** $\delta = 1; m_i = 60; \beta = 0.1; c = 5$

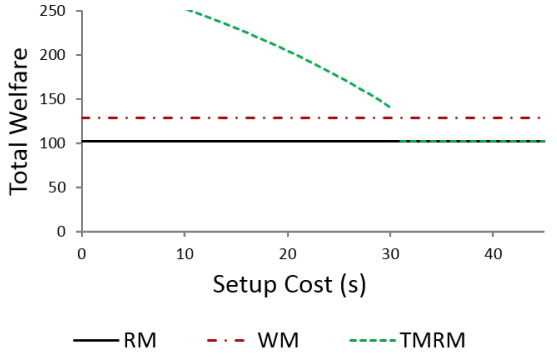
has been increasingly popular for treating minor medical conditions. Even though care may be somewhat inferior, the convenience gained by patients through telemedicine is still significant.

Figures **????** and **????** compare the total welfare for a revenue-maximizing specialist before telemedicine (RM), the total welfare for a welfare-maximizing specialist before telemedicine (WM), and the total welfare for a revenue-maximizing specialist after telemedicine (TMRM) as functions of the relative benefit of telemedicine visits, m_t/m_i (Figure **????**), and the setup cost of telemedicine, s (Figure **????**).

Figures **????** and **????** compare the total patient surplus for a revenue-maximizing specialist

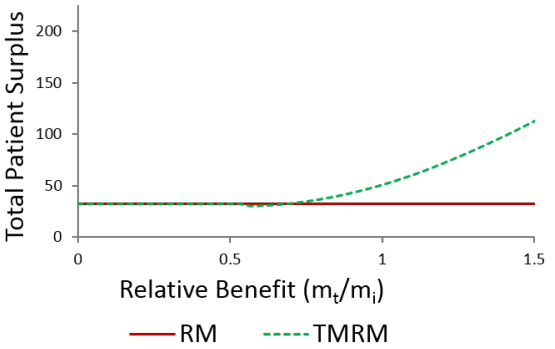


(a) Total Welfare vs. Relative Benefit of Telemedicine Visits $(\boldsymbol{m_t/m_i})$: $\delta = 1; m_i = 60; t_0 = 10; \beta = 0.1; s = 10; \alpha = 0.75; c = 5$

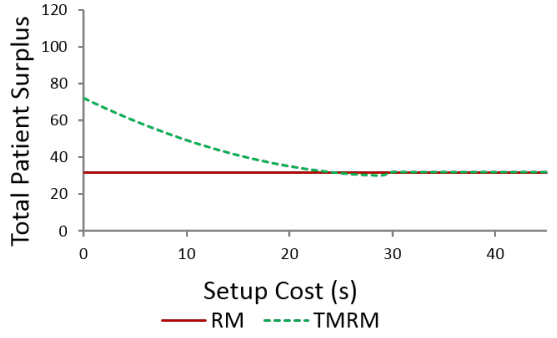


(b) Total Welfare vs. Setup Cost $(\boldsymbol{s})$: $\delta = 1; m_i = 60; m_t = 60; \beta = 0.1; t_0 = 10; \alpha = 0.75; c = 5$

Figure 13 **Analysis of total patient welfare under a welfare-maximizing specialist before the introduction of telemedicine (WM) and a revenue-maximizing specialist before (RM) and after the introduction of telemedicine (TMRM)**



(a) Total Utility vs. Relative Benefit of Telemedicine Visits $(\boldsymbol{m_t/m_i})$: $\delta = 1; m_i = 60; \alpha = 0.75; \beta = 0.1; t_0 = 10; s = 10; c = 5$



(b) Total Utility vs. Setup Cost $(\boldsymbol{s})$: $\delta = 1; m_i = 60; m_t = 60; \beta = 0.1; t_0 = 10; \alpha = 0.75; c = 5$

Figure 14 **Analysis of total patient surplus under a revenue-maximizing specialist before the introduction of telemedicine (RM) and after the introduction of telemedicine (TMRM)**

before telemedicine (RM) and after telemedicine (TMRM) as functions of the relative benefit of telemedicine visits, m_t/m_i (Figure **????**), and the setup cost of telemedicine, s (Figure **????**). In both figures, when the specialist does not offer the telemedicine mode the total patient surplus is the same ($m_t/m_i \leq 0.53$ in Figure **????** and $s \geq 30$ in Figure **????**), and RM and TMRM overlap.

Figure **??** shows how patient utility varies with distance before (Before TM) and after telemedicine (After TM). The point where the slope changes for the After TM line denotes the

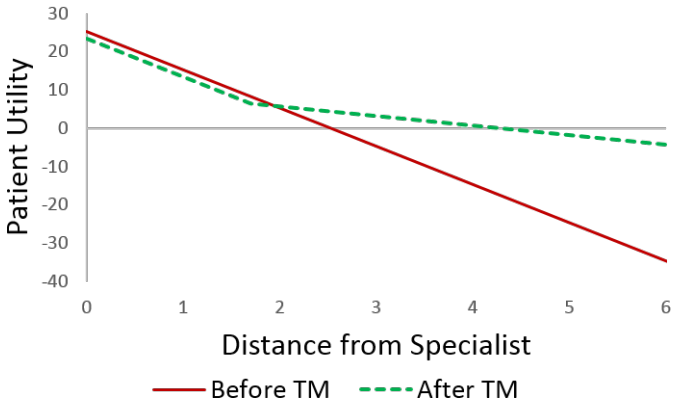


Figure 15 **Analyzing patient utility with distance from the specialist:** $\delta = 1; m_i = 60; \beta = 0.1; s = 10; c = 5; m_t = 60; t_0 = 10; s = 10; \alpha = 0.75$

threshold where patients start preferring the telemedicine mode to the in-person mode. From Figure **??**, we can observe that the utility for those patients located close to a specialist actually falls (although only marginally) after the introduction of telemedicine. Even though prices fall after the introduction of telemedicine, the higher service quality cost and the higher congestion cost reduce the utility for these patients. Telemedicine is thus beneficial only for patients located at some distance from the specialist.