The Process Implications of using Telemedicine for Chronically Ill Patients: Analyzing Key Consequences for Patients and Medical Specialists

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 very heterogeneous set of patients. Such heterogeneity arises because of difference in patients’ medical conditions as well as the travel burden each of them faces when trying to reach the clinic. In this paper, we investigate the impact of travel burden on the strategic behavior of medical specialists in terms of their operating decisions. In addressing this problem, we expand current results in queuing theory related to the service speed-quality trade-off for both revenue-maximizing and welfare-maximizing servers in the context of customers with a general and heterogeneous utility function. We find that as the relative travel burden increases for the patients, it also has a negative effect on the productivity of the specialist and his expected income. Comparing the optimal service characteristics of revenue-maximizing and welfare-maximizing specialists we see that the former will overall serve a smaller patient population, will have shorter waiting times, and will operate at a lower utilization. We also analyze the impact of the newly introduced telemedicine technology on patient utility and the specialists’ operating decisions. We prove that with the introduction of telemedicine the revenue-maximizing service rate moves closer to the socially optimal one. While the enhanced access to specialist care increases the overall social welfare, we explain why some patients, unexpectedly, will be even worse-off with the introduction of this technology. Our analytical results lead to some important policy implications for facilitating the further deployment of telemedicine in the care of chronically ill patients.

Revised: August 27, 2015  
***Working Paper***
1. Introduction

Unlike product quality, service quality largely depends on perception. Service providers have employed different mechanisms to signal quality, ranging from price (Chen et al. 1994) to service environment or ambiance (Brady and Cronin Jr. 2001). Server-customer interaction adds another vital dimension. In a healthcare setting, this interaction between the service provider (specialist) and the customer (patient), or more technically the consultation time, gains further importance. Even though a specialist might be able to cover all possible clinical aspects in a short consultation, if the specialist sees patients too fast, patients might not be satisfied with the specialist’s effort; see Gross et al. (1998) and Kong et al. (2007).

Medical specialists, just like a service provider in almost all settings, faces a trade-off: While a customer would like more time with the server, spending more time with each customer reduces the number of customers that can be served in a given time, and the slower service rate delays other customers waiting to be served. It is well known that longer waiting times reduce customer satisfaction (see, for example, Stone (2012)), and more so if the waiting time is significant when compared to the actual service time. But it is only common in medical settings that patients have to wait for hours for only a few minutes of interaction.

This paper is motivated by our long-term study of the service requirements of chronically ill patients who must visit their providers periodically to manage their condition. Chronic conditions represent a huge burden on the overall modern health care system. In an analysis of the causes of mortality and disability in the United States, chronic conditions were found to be a major factor by Jemal et al. (2005). About 140 million Americans have at least one chronic condition and these patients account for 75% of the total U.S. health expenditures as reported in the chartbook prepared by the Partnership for Solutions, see Anderson et al. (2004). By 2030 it is anticipated that nearly half of all Americans will have at least one chronic condition.

Dussault and Franceschini (2006) point out that most specialists are concentrated in urban areas, and there are few or even none in remote or rural areas. Thus while patients located in urban areas are closer to specialists, many rural patients travel long distances to visit a specialist. In fact, one-way travel time of 4 hours is not uncommon for a visit duration of just 30 minutes for patients from rural areas, as mentioned in Dorsey et al. (2013). Hence, patients are not homogeneous in the utility they derive from treatment by specialists. Besides distance, there are also other reasons for heterogeneity among patients such as their health condition.

The analysis of service systems when strategic customers decide whether to seek service based on the service quality has been studied extensively; see Naor (1969), Edelson and Hilderbrand (1975),
Hassin and Haviv (2003), Hassin and Roet-Green (2012), and Debo et al. (2012), among others. Anand et al. (2011) 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 as explained above.

Our main results in the first part of this paper are as follows. We find that the revenue-maximizing specialist works more slowly and treats fewer patients than a welfare-maximizing specialist when the patients are heterogeneous. As a result, the utilization of a welfare-maximizing specialist is higher than that of a revenue-maximizing specialist. 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 the waiting time for all patients.

One way to reduce the travel burden for patients is by means of remote medicine or telemedicine. Telemedicine is the use of electronic communication to improve patient health. Telemedicine can take a number of 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; see Schwamm et al. (2004)) or it can involve direct communication between the physician and the patient.

The direct virtual communication with a specialist holds a good deal of promise, as it can offer specialist access and timely treatment to those located far from a medical practitioner or facility. In the second part of this paper, we focus on consultations related to chronic conditions, which often are required multiple times a year to manage the underlying disease. Without these periodic visits the medical condition deteriorates. We extend the basic model that we use to study the quality-speed trade-off to include the case when patients also have the option to seek treatment via telemedicine.

For many medical conditions, it is clinically feasible to substitute virtual visits via a telemedicine facility for some in-person visits to a medical practitioner. This approach 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 Bashshur et al. (2000), Hersh et al. (2006), Kennedy et al. (2009), and Dorsey et al. (2013) for various applications, and Hersh et al. (2001), Hailey et al. (2002), and Monnier et al. (2003) for reviews of telemedicine’s benefits.

Apart from clinical efficacy and efficiency, there are several operational issues that are of interest: Given the advantages and limitations of telemedicine, what clinical fields are likely to benefit most?
What will the economic impact of telemedicine be? What kind of influence will telemedicine have on specialists’ behavior? Given the severe shortage of specialists even at the current level of coverage (see Ziller (2003)), 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 utility (surplus) for each patient by charging higher fees? Will the prices be different before and after telemedicine? How will service levels (utilization) change? In short, how will telemedicine affect specialists and providers and what will the subsequent impact on patients and on social welfare be?

By incorporating a factor for clinical feasibility and efficiency into our analysis, we look at the operational, managerial, and economic issues related to the implementation of telemedicine. With the introduction of telemedicine, we find that the utilization of a revenue-maximizing specialist goes up, his service rate increases, and he sees more patients. In addition, we establish a relatively simple necessary condition to identify when telemedicine is economically feasible. According to this result there are three important parameters that determine the feasibility of telemedicine: the perceived “quality” difference between telemedicine and in-person visits, the travel cost (or disutility) because of the distance the patients have to travel without telemedicine, and the technological and maintenance costs for a patient to receive telemedicine treatment.

We also present a necessary condition for the introduction of telemedicine to increase the total welfare. Based on this necessary condition 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; see Hubble et al. (1993), Samii et al. (2006). Also, many Parkinson’s patients have a relatively high travel cost, in part because neurologists are usually located in large urban areas; see Dorsey et al. (2013). Therefore, treatment of Parkinson’s patients is highly suitable for implementing telemedicine, assuming the technological cost is not excessive.

The technical contributions of our paper are two-fold. First, we extend the literature on the analysis of queuing systems with strategic customers. Due to customer heterogeneity the standard results from the literature cannot be used directly. We 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 a revenue maximizing and a welfare maximizing service provider. For the analysis of the specialist’s decisions when telemedicine is introduced we use the concept of non-atomic games, reviewed in Section 1.2. We establish the Nash equilibrium in our queuing model and then we use this result to find the optimal decisions for the specialist.
1.1. Related Literature

Anand et al. (2011) model customer service time intensity and study the interaction between service value and speed. They consider a single server, serving homogeneous customers. They also consider the quality effect of a given service rate and incorporate 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 (quality effect), the waiting cost incurred, and the price to be paid.

Heterogeneity in service values for customers has been considered in the queuing literature beginning with Littlechild (1975) and later Larsen (1998). But the literature largely assumes that the quality of service is not affected by the service rate, and hence the interaction was not studied until Anand et al. (2011). Mendelson (1985) considers optimizing over price and capacity (service rate) with the cost of capacity being convex in service rate. But he does not consider the cost of service to be part of a customer’s utility function. To our knowledge, service interaction for the case of heterogeneous customers has not been studied in the literature. We refer the reader to Anand et al. (2011) and Hassin and Haviv (2003) for an extensive review of the literature. Naor (1969), Larsen (1998), and Shone et al. (2013) consider and compare revenue-maximizing and welfare-maximizing servers. Also, Wright (2007) demonstrates how different objectives for a specialist may arise in a health care setting.

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 Jennett et al. (2003), Dávalos et al. (2009) and Whitten et al. (2002). 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 the impact with a service model that considers the main trade-offs that arise with the introduction of telemedicine.

1.2. Non-cooperative non-atomic games

In this section we provide a general description of non-cooperative non-atomic games, which we use to study patient choices. 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. To capture these features we next define a general form for a game.

We follow the terminology in Schmeidler (1973). 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 $\sigma$-algebra, endowed with an absolutely continuous probability measure $\mu$ with respect to the Lebesgue measure. Measure $\mu$ 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 $A = \{1, \ldots, n\}$ denote the action set for patients. For our purposes it is enough to consider a discrete $A$ and an action set that is not dependent on $x$. Denote

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

Let $g_x \in P$ denote a distribution on $A$ 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 $\Psi_a(\Gamma, G^{-1}(x), x)$ 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 $\Gamma$. (In our context $\Gamma$ will be 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 $\mu$-integrable. In this paper we restrict our attention to T-strategies as in [Schmeidler (1973)].

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

$$\sum_{a \in A} g_x(a)\Psi_a(\Gamma, G^{-1}(x), x) \geq \sum_{a \in A} p_x(a)\Psi_a(\Gamma, G^{-1}(x), x) \quad \mu\text{-a.s. for } x \in \mathbb{R}^+ \text{ and for any } p_x \in P.$$  \hfill (2)

Inequality in (2) implies that under the strategy profile $G$ none of the patients is better off by deviating from her 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; see [Schmeidler (1973), Rath (1992) and Khan and Sun (2002)]. In [Schmeidler (1973)], 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. Therefore, unlike in [Schmeidler (1973), Rath (1992) and Khan and Sun (2002)] there might be multiple patients with the same index. Therefore, it is not clear how $g_x$ should be interpreted. However, we will show that in our setting, we can find a pure equilibrium strategy where patients with the same index follow the same strategy (in a.s. sense). In general, one can consider $g_x$ for a mixed strategy to be the distribution on patients with index $x$ on how to choose each available action. In addition, because we assume that $\mu$ 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 payoff to other players do not change. Hence the interpretation of $g_x$ is not crucial for the applications we focus on.
2. **Effect of patient heterogeneity on specialist interaction**

In this section we consider a basic model without the telemedicine component, before turning to the effect of telemedicine on patient and provider behavior in Section 3.

### 2.1. 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 the price he charges each patient to maximize a certain objective (whose details are discussed below). In response to these choices, patients decide to seek service based on their expected net utility. Each patient’s utility comprises a reward from seeking treatment, waiting costs, quality costs, and payments.

Typically, a specialist has patients coming from various places. The net utility of a patient from seeking service thus depends on the distance to be traveled to see the specialist. Specifically, we model the patient utility from seeking treatment as a function of the patient’s distance from the specialist. We are mainly interested in this setting due to our ultimate goal to understand the effect of telemedicine on patient and provider choices (see Section 3). One of the most significant benefits of telemedicine is that it obviates the need to travel for some consultations with the specialist.

Assume that the potential patients are indexed by $x$, their distance from the specialist. Also assume that each patient gets a constant reward $m$ per visit to the specialist. Let $t : \mathbb{R}_+ \rightarrow \mathbb{R}_+$, where $t(x)$ denotes the travel burden for a patient located at a distance $x$ from the specialist. 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 $m - t(x)$. We also assume that $t$ is strictly increasing and thus invertible. Without loss of generality we assume that $t(0) \geq 0$. 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 the current health status or individual perceptions. (This assumption can be relaxed easily, see Remark 3 below.) Let $f$ denote the density function for the distribution of the distance of the patients from the specialist and $F$ denote the associated cdf.

Let $M_v = m - t(0)$, for $m < \infty$. We assume that $M_v > 0$; that is, at least some of the patients would potentially seek treatment. Let $X_v = t^{-1}(m)$ denote the farthest distance of the patient whose benefit from treatment net travel burden is zero, hence $m - t(X_v) = 0$. We note that patients whose distance is more than $X_v$ would not seek treatment no matter how the other parameters are chosen. We also assume throughout that $F : [0, X_v] \rightarrow [0, 1]$ is one-to-one and $f > 0$ (obviously the latter implies the former).

Each patient is assumed to need repeated visits to a specialist. We normalize this rate to 1 per unit time without loss of generality since we can alter the time unit if necessary. But we assume
implicitly that all patients need to see the specialist at the same rate. This should hold true for a population of patients suffering from the same chronic disease. Also, 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.

Given that the specialist chooses a service rate $\mu$ and price $p$, let the arrival rate of patient visits to the specialist be $\lambda$. The arrival rate is obtained from the number of patients who decide to seek treatment based on their utility in equilibrium, whose specifics are discussed below. Let $\mathbb{R}$ denote the real numbers and $\mathbb{R}_+ = [0, \infty)$. 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 $\lambda$, $p$, and $\mu$ is given by:

$$\Psi(\lambda, p, \mu, x) = m - t(x) - Q(\mu) - \beta p - c\mathbb{E}W(\mu, \lambda).$$

Here $Q : \mathbb{R}_+ \to \mathbb{R}_+$ denotes the cost or disutility to the patient as a function of the service rate $\mu (\geq 0)$. (We assume without loss of generality that $Q \geq 0$ for simplicity.) The price charged by the specialist is denoted by $p$, and we assume that a patient pays $\beta p$ for each visit, for $0 < \beta \leq 1$. The total arrival rate of patient visits per unit time is denoted by $\lambda$, and the expected time spent by the patient in the system is given by $\mathbb{E}W(\mu, \lambda)$, as a function of the service rate $\mu$ and the arrival rate $\lambda$. To model the fact that patients value their time, patients incur an opportunity cost per unit time they spend in the system, which we denote by the parameter $c$.

Since patients typically are covered by insurance, most of the cost is borne by their insurance. Thus, if the specialist charges a price $p$ for the services rendered, the patient pays $\beta p$, where $\beta$ 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.

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 [Green and Savin (2008)] and [Anand et al. (2011)] for similar assumptions), thus resulting in an M/M/1 queuing model, an assumption that helps provide analytical tractability. With the M/M/1 assumption, $\mathbb{E}W(\mu, \lambda) = \frac{1}{\mu - \lambda}$, so the net utility for the patient is then given by

$$\Psi(\lambda, p, \mu, x) = m - t(x) - Q(\mu) - \beta p - \frac{c}{\mu - \lambda}. \quad (3)$$

Our results can readily be extended to cover other inter-arrival and service time distributions using heavy traffic approximation methods given by [Kingman (1961)]. Although we do not explore them here, approximation methods discussed in [Shanthikumar et al. (2007)] can be used for similar systems. We also assume that patients do not renege, that is, our model does not allow no-shows.
Though the exact waiting time is known when patients call for an appointment with the specialist, multiple visits are required to manage a chronic condition as explained before. Also, patients usually stick with the same provider for care continuity. Although they may switch to another provider in the long run, it is not very plausible to switch to another provider for every visit. Hence patients do not have sufficient information on how many other patients will be ahead of them for subsequent appointments; therefore we model the queues as unobservable and their decision based on long run averages, similar to Edelson and Hilderbrand (1975), Chen and Frank (2004), and Anand et al. (2011). Even though patients do not observe the actual queue length, they are assumed to have an expectation of the time they would have to spend at the clinic or the specialist’s office based on past visits and repeated interactions with the specialist. Because telemedicine is especially used for chronic patients, repeated visits by a patient to the same specialist are commonplace; see Starfield et al. (2003) and Wennberg (2002). Hence, we assume patients make decisions depending upon the expected waiting time. Similarly, we assume that patients know the average time they will spend with the specialist from their past experience. We refer to Chapter 3 of Hassin and Haviv (2003) for an excellent summary on unobservable queues and their applications.

Remark 1. Our utility model builds on extant literature. While Littlechild (1975) and Chen and Frank (2004) consider customer heterogeneity, they do not consider the quality cost of service, \( Q(\mu) \). Anand et al. (2011) consider \( Q(\mu) \) to be linear but assume the benefit to be homogeneous. Specifically, they take \( m - t(x) - Q(\mu) = (V_b + \alpha \mu_b - \alpha \mu)^+ \), where \( V_b, \alpha, \) and \( \mu_b \) are all non-negative constants. (It can be shown that the service provider’s optimal decisions do not change if \( t(x) = 0 \) \( \forall x \); see Section 2.4.) By considering the quality cost, \( Q \), as well as customer heterogeneity, we extend the growing literature on service operations. We are especially interested in the cost for patients associated with traveling to a specialist’s office, which requires considering customer heterogeneity.

2.1.1. Patients’ utility and equilibrium: As mentioned above, patients have two choices, either to seek treatment or not, based on their 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. Thus, 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 obtain a positive net utility. Let \( \lambda(\cdot, \cdot) : \mathbb{R}_+^2 \rightarrow \mathbb{R} \), where \( \lambda(p, \mu) \) is the equilibrium arrival rate of such rational self-interested patients per unit time for fixed \( p \geq 0 \).
and $\mu \geq 0$. Since the decision is taken prior to joining the queue, the customer does not balk after reaching the specialist’s facility.

Following the definitions in Section 1.2, the action set for the patients is given by $A = \{i, o\}$, where “$i$” denotes (in-person) treatment and “$o$” denotes no treatment. Because we assume that a patient not seeking treatment receives a reward equal to zero, a T-strategy $G = \{g_x : x \geq 0\}$ is an equilibrium if (2) holds with $\Psi$ defined as in (3). The arrival rate for a given T-strategy is then given by

$$\lambda(p, \mu) = \Lambda \int_0^{X_m} g_x(i)f(x)dx,$$

since each patient is assumed to seek treatment at rate 1, if the patient chooses to seek treatment. We also note that because of this normalization $\lambda(p, \mu) \leq \Lambda$, for all $p \geq 0$ and $\mu \geq 0$.

Because of the special structure of the utility function $\Psi$, it is not difficult to see that for a given $(\mu, p)$, an equilibrium (in the sense defined in (2)) has to have the following structure: There exists a threshold, $x^*$, such that patients whose distance from the specialist $x$ satisfies $x \leq x^*$ seek treatment, $g_x(i) = 1$ for $x \leq x^*$, and other patients do not seek treatment, $g_x(o) = 1$ for $x > x^*$. By (4), the arrival rate under this equilibrium is given by

$$\lambda(p, \mu) = \Lambda F(x^*).$$

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

$$x^* = \inf\{x \geq 0 : \Psi(\Lambda F(x), p, \mu, x) \leq 0 \text{ and } \Lambda F(x) \leq \mu\} \land X_m,$$

where the condition $\Lambda F(x) \leq \mu$ ensures stability and the inf of an empty set is taken to be 0 by convention. To explain the intuition behind (6), assume that $\Psi(\Lambda F(x^*), p, \mu, x^*) = 0$ for some $x^*$. Then the patient who is at a distance $x^*$ from seeing 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.

2.1.2. Revenue-maximizing specialist: Given patient utility function $\Psi$, when the specialist is paid on a fee-for-service basis the revenue is also directly proportional to the equilibrium arrival rate. The specialist has two decision variables that he typically has control over, price and service rate. Each specific choice of this pair automatically determines an equilibrium arrival rate.
based on the patient utility function as we explained in the previous section (see Mendelson (1985), Hassin and Haviv (2003), and Anand et al. (2011) for a similar approach in different applications).

The specialist then has the revenue function $R : \mathbb{R}_+^2 \rightarrow \mathbb{R}$ defined by

$$R(p, \mu) = p\lambda(p, \mu).$$  \hfill (7)

Note that the specialist is facing a complex and a non-intuitive trade-off. If the specialist tries to increase price then the equilibrium arrival rate falls and hence the marginal revenue may either increase or decrease. Similarly, if the specialist works faster, even though the waiting costs will be lower, increasing the utility of the patients and thus increasing the equilibrium arrival rate, it will also increase the quality cost as well and hence also decreases the equilibrium arrival rate. The opposite effects can be seen if the specialist tries to work slowly. The revenue-maximizing specialist will try to choose an optimal price and service rate to obtain the maximum revenue given by

$$R^* = \sup_{p \geq 0, \mu > 0} R(p, \mu).$$  \hfill (8)

If the payments are out-of-pocket, then modeling price as a decision variable for the specialist is a plausible approach. Even when prices are determined by CPT codes and reimbursed by the insurance provider, the specialists, through their associations, have sufficient say in determining the prices in the long run; see Hariri et al. (2007) and Wasik (2013). Besides there are also doctors who bypass the insurance system and charge the patients directly; see Sullivan (2012) and Hargreaves (2013). Also, in countries like Australia, France and Finland, medical practitioners, especially specialists, are free to set their own prices for the services they offer; see AMA (2015), Kumar et al. (2014) and Mossialos and Srivastava (2008). Hence we assume price to be one of the decision variables for the specialist. Similarly, while a minimum time with each patient is recommended, which can be accounted for in $Q$, the specialist usually has some degree of flexibility as to how much time he spends with each patient on average; see Jauhar (2014). We thus also assume the average service rate to be a decision variable for the specialist.

2.1.3. 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 Littlechild (1975). If the welfare-maximizing specialist cannot serve all patients, he would prefer to choose the patients with the highest benefits from treatment. Other parameters remaining the same, the closer the patients are located the higher the benefits received by them. 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. We next show that we can translate the problem to choosing an arrival rate instead of the maximum distance.

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

$$\lambda_x = \Lambda F(x).$$

(9)

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

$$x_{\lambda} = F^{-1}(\lambda/\Lambda) \text{ for } \lambda \in (0, \Lambda), x_{\lambda} = X_m \text{ if, } \lambda = \Lambda, \text{ and, } x_{\lambda} = 0 \text{ if } \lambda = 0.$$  

(10)

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

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

(11)

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

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

(12)

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

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

(13)

In passing we highlight the fact that by Leibniz’s rule, $V'(\lambda) = m - t(F^{-1}(\lambda/\Lambda))$ for $\lambda \in (0, \Lambda)$.

Hence by (10), 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.

Before we proceed we would like to provide more insight into the objective of a welfare maximizer. With each additional patient, the workload of the specialist goes up, so if the average service rate remains the same, the waiting time increases for all patients seeing the specialist. Each additional patient 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 [12].
2.1.4. Technical assumptions

In order to be able to obtain a solution for the revenue-maximizing and welfare-maximizing specialists we need to make certain assumptions. We list these assumptions in this section and explain why they are plausible. We make the following technical assumptions on $V$ and $Q$. Throughout we assume that the derivative of a function denotes its right derivative at the left boundary and left derivative at the right boundary of its domain. First, we assume that $m < \infty$, hence $V(\lambda) < \infty$ by (11) for all $\lambda \in [0, \Lambda]$, and $Q(\mu) < \infty$ for all $\mu \geq 0$.

Assumptions on $V$:

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

A.2 $t : [0, X_m] \rightarrow \mathbb{R}$ 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 non-decreasing rate. The assumption is true in many general cases in which patients dealing with chronic conditions are old and hence 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, $f(x) = e^{-x}$). This is because specialists tend to be located in urban areas and the population density tends to be lower in rural areas.

**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:

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

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

The proofs of the results in this section are in Appendix B.1. Note that $V(0) = 0$ and $m - t(0) > 0$. If nobody seeks service, then the total value of all completions of service is zero, and $m - t(0) > 0$ implies that, at least when all other costs are absent, there is an incentive for at least one patient to join the system. Since the travel burden increases with distance, the utility decreases with distance, and hence $V'(\lambda)$ is decreasing in the arrival rate. Since the marginal value, $V'(\lambda)$, is decreasing, $V(\lambda)$ is concave. The marginal value, $V'(\lambda)$, is also concave capturing the steep fall in marginal value as more and more patients join the system.

Assumptions on $Q$:

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

B.2 $Q : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ 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 by the specialist with the patient and thus decreases with the service rate chosen by the
specialist. Assumption B.1, which combined with Assumption B.2 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 B.2 also helps model the fact that the specialist is heavily penalized for working too fast. In addition, the convexity in Assumption B.2 helps us simplify sufficient conditions for finding the optimal values. Assumptions B.1 and B.2 also imply that

$$\lim_{\mu \to \infty} Q(\mu) = \infty. \quad (14)$$

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 B.2 implies $Q'(\mu) < \infty$, $Q''(\mu) < \infty$, for all $\mu \in \mathbb{R}^+$, which we use in the proofs of our main results.

### 2.2. Optimal decisions of a revenue-maximizing specialist

We next find the optimal actions of a revenue-maximizing specialist given patient utility function $\Psi$, defined in (3). The main technical difficulty is that in the definition of $x^*$ in (6) 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 (8) 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 (6). It turns out that it is easier to solve the optimization problem with arrival rate and 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 (7), as we show next.

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

$$p(\lambda, \mu) = \inf \left\{ p \in \mathbb{R} : m - t(F^{-1}(\lambda/\Lambda)) - Q(\mu) - \beta p - \frac{c}{\mu - \lambda} \leq 0 \right\} \lor 0, \quad (15)$$

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 $\mu$ and the arrival rate $\lambda$. 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, 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,

$$\tilde{R}(\lambda, \mu) = \begin{cases} p(\lambda, \mu) \lambda, & \text{if } \lambda \in [0, \Lambda] \text{ and } \mu > \lambda \\ 0, & \text{otherwise} \end{cases}. \quad (16)$$
The specialist’s objective is to maximize his revenue by choosing an appropriate service rate $\mu$ and arrival rate $\lambda$. We will thus be able to define the following optimization problem:

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

Before we verify that the optimization problem in (17) is equivalent to (8), 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 $\mu \leq \lambda + \frac{c}{M_v}$, then $\tilde{R}(\lambda, \mu) = 0$ for any $p \geq 0$ and $\lambda \in [0, \Lambda]$. Also, if $\lambda \geq \kappa$, then $\tilde{R}(\lambda, \mu) = 0$ for all $\mu \geq 0$.

The proofs of the results in this section are in Appendix B.2. 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 2 and (15). Hence $\kappa < \Lambda$. Lemma 2 follows from the fact that the specialist can work neither too quickly nor too slowly. For if he treats patients too quickly, then the cost of quality will be too high, because $Q(\mu)$ is decreasing in $\mu$, resulting in a negative net utility for all the patients. Similarly, if the specialist treats patients too slowly, waiting time, and hence waiting cost, will become higher, resulting in a negative net utility for all the patients. (We note that Lemma 2 is similar to the quantities $A_1(\alpha)$ and $A_2(\alpha)$ in page 43 of Anand et al. (2011), except that our result is valid for a more general model.)

Mathematically, Lemma 2 helps in finding a lower and upper bound for the optimal service rate. We will thus be able to write the optimization problem in (17) using Lemma 2 as

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

The boundary conditions in (18) follow from Lemma 2. We have the following lemma to prove that the optimization problem in (18) is equivalent to (8):

**Lemma 3.** For $R$ defined as in (7) and $\tilde{R}$ as in (16), 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 $R^* = R^*$. 
We can now concentrate on finding the optimal service rate for the specialist and the optimal \( a \) that the specialist must choose to admit (the arrival rate). The optimal price can then be determined using (15).

**Solution to the optimization problem:** We solve the optimization problem in (18) in two sequential steps. First, we solve the following optimization problem for fixed \( \lambda \geq 0 \), i.e. we determine

\[
\tilde{\mathcal{R}}^*(\lambda) = \sup_{\lambda^+ \leq \mu} \tilde{\mathcal{R}}(\lambda, \mu).
\]  

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

\[
\gamma(\lambda) = \begin{cases} 
\mu : Q'(\mu) - \frac{c}{(\mu - \lambda)^2} = 0 \text{ and } \mu > \lambda 
\end{cases}.
\]  

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, \( 0 < \gamma'(\lambda) = \frac{2c}{2c + (\gamma(\lambda) - \lambda)^2 Q''(\gamma(\lambda))} \leq 1 \)

Next we show that \( \gamma \) gives the optimal service rate.

**Lemma 5.** Given \( \lambda \in (0, \kappa) \), if there exists \( \mu(> 0) \) such that \( \tilde{\mathcal{R}}(\lambda, \mu) > 0 \), then \( \gamma(\lambda) \) gives the optimal solution for (21); that is, \( \tilde{\mathcal{R}}^*(\lambda) = \tilde{\mathcal{R}}(\lambda, \gamma(\lambda)) \). If \( \lambda \geq \kappa \), or \( \lambda = 0 \), or \( p(\lambda, \mu) = 0 \) for all \( \mu \geq 0 \), then \( \tilde{\mathcal{R}}^*(\lambda) = 0 \).

By (22) and Lemmas 4 and 5 the optimal service rate is increasing in the arrival rate. More interestingly the derivative, \( \gamma' \), is bounded by 1. In other words, in response to an increase in the arrival rate, the optimal capacity does not increase by the same rate. The implication is that because the service rate does not increase as much, congestion costs are likely to increase if the arrival rate increases. Thus, if the arrival rate is exogenous and increases, though the specialist will respond with an increase in his service rate, the average waiting time for all patients will increase owing to increased congestion.

We next solve the optimization problem for \( \lambda \) to complete the optimization problem in (17). We do so by substituting the optimal service rate for a given \( \lambda \) from Lemma 5. Assume that \( \lambda \in [0, \Lambda] \) satisfies the following equation (we show that the solution is unique if a solution exists):

\[
V'(\lambda) + \lambda V''(\lambda) - Q(\gamma(\lambda)) - \frac{c}{\gamma(\lambda) - \lambda} - \frac{c\lambda}{(\gamma(\lambda) - \lambda)^2} = 0.
\]  

\( (23) \)
Theorem 1. If there exists $(\lambda, \mu)$ such that $\tilde{R}(\lambda, \mu) > 0$, then there exists a unique optimal $\lambda^*$; that is, $\tilde{R}(\lambda^*, \gamma(\lambda^*)) = \tilde{R}^* > 0$ and $\lambda^* = \lambda$. If no such $(\lambda, \mu)$ exists, $\tilde{R}^* = R^* = 0$.

Given functional forms for the utility functions, the optimal service rate can thus be determined using Lemma 5. Finally, using (15), the price can be determined as well, solving the optimization problem (8) by Lemma 3.

We next analyze the effect of the travel burden on the equilibrium arrival rate and the optimal service rate for the revenue-maximizing specialist. The following result follows from Theorem 1 and (11).

Proposition 1. Consider two travel burden functions $t_1$ and $t_2$ such that $t_1(x) = a$ and $t_2(x) = t(x) + a$ for some constant $a \geq 0$, for all $x \geq 0$. Let $(\lambda^*_i, \mu^*_i)$ 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 $i = 1, 2$ under the assumptions of Lemma 1 and assume that Assumptions B.1 and B.2 hold. The following results hold:

(i) $\lambda^*_1 \geq \lambda^*_2$,

(ii) If in addition $\lambda^*_1 > 0$ and $\lambda^*_2 > 0$, then $\mu^*_1 \geq \mu^*_2$.

The implication of this result is that if the distance cost or traveling cost reduces the utility of seeking treatment and hence fewer patients are interested in reaching out to a specialist, then the revenue-maximizing specialist sees fewer patients. The second part of the proposition implies that quality costs decline as the service rate decreases. 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 the patients. In other words, because the patients travel from far-off places or go through great difficulty in coming to a specialist, the specialist tries to spend more time with them to make them feel better, thereby increasing their net utility.

Remark 2. The model in this section can easily be used to study the scenario when prices are exogenous. From (3), (5) and (6), the arrival rate decreases if the distance cost or traveling cost increases. From (22), then the service rate will decrease as well, hinting that the specialist will try to compensate for the travel burden.

Remark 3. The model in this section can be extended to account for heterogeneous treatment utilities (i.e. random $m$) by redefining $f$, the distribution for the utility, to capture the treatment utility and the traveling burden. In addition it can be shown with some additional work that Proposition 1 still holds if we take $t_1(x) = 0 \leq t_2(x)$ for all $x \geq 0$. The former case ($t_1(x) = 0$) corresponds to the model in Anand et al. (2011).
2.3. Optimal decisions of a welfare-maximizing specialist

In Section 2.2, we considered the specialist and the patients to be separate entities focused on their own self-interest. In this section, we will consider the specialist and the patients as a single combined entity and solve the optimization problem in (13) for socially optimal decisions. The technical details are similar to those in Section 2.2, 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):

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

**Theorem 2.** If there exists \((\lambda, \mu)\) such that \( U(\lambda, \mu) > 0 \), then there exists a unique optimal \( \lambda^*_s \); 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 \((\lambda, \mu)\) exists, \( U^* = 0 \).

Comparing Theorems 1 and 2, if the number of patients treated per unit time by 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 to the specialists are exogenous, then both the revenue-maximizing and the welfare-maximizing specialists spend equal amounts of time with the patients on average.

2.4. Comparing decisions for the two different objectives

Now that we have the optimal arrival rates and service rates for both types of specialist, we can compare them.

**Proposition 2 (Comparison of the two specialists).** 1. 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 \( \tilde{R}^* > 0 \).

2. 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 \( \lambda^* > 0 \).

The proof follows from Theorems 1 and 2, in a way similar to the proof of Proposition 1. From Proposition 2, a revenue-maximizing specialist sees fewer patients per unit time than a welfare-maximizing specialist sees per unit time. 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 charged by the specialist and the travel burden. Hence, the total welfare is lower. We note here that this part of the proposition is similar to what Mendelson
found, and our proof follows similar lines as well, but because of the addition of the quality effect the extension is not straightforward. The first part of the result is obviously 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, from Lemma 5, $\gamma'(\lambda) \leq 1$, so even though the service rate increases and the specialist sees more patients in the welfare-maximizing case, the congestion cost is not lower than it is in the case of a revenue-maximizing specialist.

**Remark 4.** As mentioned in the introduction, besides distance, there are also other reasons for heterogeneity among patients such as their health condition. We note here that Lemma 1 merely states sufficient conditions for $t$ and $F$. As long as the conditions stated in the lemma for $V$ are valid, our results can be used to model other cases of heterogeneity as well. We only need the distribution of the utility to satisfy the conditions expressed in Lemma 1 for $V$.

### 3. Role of telemedicine

In this section, we model the implications of telemedicine for utility-maximizing patients and a revenue-maximizing specialist. 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 may be done via telemedicine.

We assume that the specialist offers two modes of treatment, telemedicine and in-person. A patient thus has three choices: (1) in-person mode; (2) telemedicine mode; and (3) no treatment. A patient’s choice will depend on the respective utilities of each option to her. Because a patient who chooses the telemedicine mode still has to visit the specialist in person for a certain fraction of her consultations with the specialist, 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.

Similar to the model in Section 2.1, 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 we assume $t$ is strictly increasing.)
Term $t(0)$, being a constant, can be adjusted in the factor $m_i$ and $m_t$, so we take $t(0) = 0$ without loss of generality. Let $M = \max\{t^{-1}(m_i), t^{-1}(\frac{\alpha}{1-\alpha}m_i + m_i)\}$, where $\alpha$ is the fraction of clinical visits possible via telemedicine and the remaining fraction, $(1-\alpha)$, involves visiting the specialist in person) denote the maximum distance a patient might seek treatment from the specialist. We again use $f$ and $F$ to denote the pdf and the cdf for the distribution of the patients on $[0, M]$, respectively, and that they satisfy the conditions in Lemma 1 on $[0, M]$.

Next in Section 3.1 we introduce a model where 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 3.2 we consider the case where the specialist can choose different service rates and prices. For the latter model, the equilibrium behavior of patients is much more complex, so we make additional assumptions and offer numerical results after we present a method to identify the equilibrium.

3.1. 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 this section, we assume that the specialist does not differentiate between telemedicine and in-person visits in terms of the service rate and the price he charges each patient, but instead chooses a service rate $\mu$ and price $p$ for all patients. From the specialist’s perspective, the patient’s mode of treatment does not really matter. However, from a patient perspective, the patient’s utility is dependent upon the mode of treatment due to different travel burdens.

If a patient chooses in-person treatment, her utility can be captured in a way similar to that in Section ?? However, we find it more convenient to index the patients by their distance from the specialist. For a given total arrival rate $\lambda$, price $p$, service rate $\mu$ and the distance of the patient, $x$, we define the patient utility $\Psi_i : [0, \Lambda] \times \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R}$ for in-person mode by

$$\Psi_i(\lambda, p, \mu, x) = m_i - t(x) - Q(\mu) - \beta p - \frac{c}{\mu - \lambda}. \tag{24}$$

The main difference between $\Psi_i$ and $\Psi$ in (3) (besides the difference in indexing) is that $\lambda$ is the total arrival rate, including those who chose the telemedicine mode. On the other hand, if the patient located at distance $x$ from the specialist chooses the telemedicine mode, the patient utility $\Psi_t : [0, \Lambda] \times \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R}$ in our model is given by

$$\Psi_t(\lambda, p, \mu, x) = \alpha m_t + (1-\alpha)(m_i - t(x)) - s - Q(\mu) - \beta p - \frac{c}{\mu - \lambda}. \tag{25}$$
As discussed earlier, based on our research experience with Parkinson’s patients, not all the care provided in the office can be delivered by telemedicine sessions. A certain portion of the annual care of visits still require an in-person visit with a specialist because they may involve certain physical exams, tests or procedures that cannot be done remotely. Accordingly, for telemedicine patients we assume that the fraction of clinical visits possible via telemedicine is given by \( \alpha \in (0, 1) \) and the remaining fraction, \( (1 - \alpha) \), involves visiting the specialist in person. Also, \( s \in \mathbb{R}_+ \) denotes the amortized cost for 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 Polisena et al. (2009). It also includes the operating cost and any cost related to on-site technical support per visit. The other terms are similar to those in Section 2.

3.1.1. Equilibrium for a specialist offering both modes

In this section, we establish the equilibrium when patients have a choice between in-person and telemedicine modes based on the definition in Section 1.2. 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 \( \mathcal{A} = \{i, t, o\} \). The reward functions are given by \( \Psi_i \) and \( \Psi_t \), defined in (24) and (25), for actions \( i \) and \( t \). Recall that we assume a patient gets zero reward if she does not seek treatment. The equilibrium definition follows (2) with these reward functions, assuming patients are indexed by their distance \( x \) from the specialist.

Given a T-strategy \( G = \{g_x : x \geq 0\} \), the arrival rates \( \lambda_{in} \) for in-person mode patients and \( \lambda_{tm} \) for telemedicine mode patients are given respectively as follows:

\[
\lambda_{in} = \Lambda \int_0^\infty g_x(i) f(x) dx \quad \text{and} \quad \lambda_{tm} = \Lambda \int_0^\infty g_x(t) f(x) dx,
\]

where \( \Lambda > 0 \) denotes the total arrival rate if all patients seek treatment. Let \( \lambda_t \) denote the arrival rate of telemedicine visits and \( \lambda_i \) 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,

\[
\lambda_i = \lambda_{in} + (1 - \alpha) \lambda_{tm} \quad \text{and} \quad \lambda_t = \alpha \lambda_{tm}.
\]

We set \( \lambda = (\lambda_i, \lambda_t) \).

We next establish the equilibrium for a given \( p \) and \( \mu \). We start with the following elementary result:

**Lemma 6.** Fix \( \lambda, p \) and \( \mu \in \mathbb{R}_+ \). If, for a patient located at distance \( x_1 \), \( \Psi_i(\lambda, p, \mu, x_1) \geq \Psi_i(\lambda, p, \mu, x_1) \), then \( \Psi_i(\lambda, p, \mu, x) > \Psi_i(\lambda, p, \mu, x) \) for all \( x < x_1 \). Similarly, if, for a patient located at distance \( x_1 \), \( \Psi_i(\lambda, p, \mu, x_1) \leq \Psi_i(\lambda, p, \mu, x_1) \), then \( \Psi_i(\lambda, p, \mu, x) < \Psi_i(\lambda, p, \mu, x) \) for all \( x > x_1 \). Also, if \( \Psi_t(\lambda, p, \mu, x_1) < 0 \), then \( \Psi_t(\lambda, p, \mu, x) < 0 \) for all \( x > x_1 \).
The proof immediately follows from the definitions of $\Psi_t$ and $\Psi_i$ in (24) and (25) and the fact that $t$ is strictly increasing. Based on this result we can deduce that for a T-strategy $G$ 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. Therefore, $g_x(i) = 1$ for $0 \leq x \leq x_0$, $g_x(t) = 1$ for $x \in (x_0, x_1]$, and $g_x(o) = 1$ for $x > x_1$.

From (24) and (25), the difference between the utilities obtained from choosing the in-person mode and the telemedicine mode for a patient at distance $x$ is given by

$$\alpha (m_i - t(x) - m_t) + s.$$  \hfill (28)

Let $x_{ID}$ be defined as follows:

$$x_{ID} = \inf \{ x \geq 0 : \alpha (m_i - t(x) - m_t) + s \leq 0 \}. \hfill (29)$$

Thus, $x_{ID}$ denotes the location of the patient who is indifferent between the two modes of treatment. From here on we assume that $x_{ID} > 0$; that is, at least some patients prefer the in-person mode. Otherwise, none of the patients prefer the in-person mode and this case reduces to a specialist offering only telemedicine mode (see Section 3.1.2). Let $D : \mathbb{R}_+^2 \to \mathbb{R}_+$ be defined as follows

$$D(p, \mu) = \begin{cases} 
\inf \{ x \geq 0 : k_1(p, \mu, \Lambda F(x)) - t(x) \leq 0 \text{ and } \mu \geq \Lambda F(x) \}, & \text{if } k_1(p, \mu, 0) \geq 0, \\
0, & \text{otherwise} 
\end{cases}, \hfill (30)$$

where $k_1(p, \mu, \ell) = m_i - Q(\mu) - \beta p - \frac{c}{\mu - \ell}$. 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 $k_2(p, \mu, \ell) = \alpha m_i + (1 - \alpha) m_i - s - Q(\mu) - \beta p - \frac{c}{\mu - \ell}$, and define $x_{TM} : \mathbb{R}_+^2 \to \mathbb{R}_+$ by

$$x_{TM}(p, \mu) = \begin{cases} 
\inf \{ x \geq 0 : k_2(p, \mu, \Lambda F(x)) - (1 - \alpha) t(x) \leq 0 \text{ and } \mu > \Lambda F(x) \}, & \text{if } k_2(p, \mu, 0) \geq 0, \\
0, & \text{otherwise} 
\end{cases}. \hfill (31)$$

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. We have the following result.

**Theorem 3.** For fixed $p \geq 0$ and $\mu > 0$,

i) if $D(p, \mu) \leq x_{ID}$, then the unique 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 \( x > x_{TM}(p, \mu) \).

The proofs of the results in this section are in Appendix C. Based on Theorem 3, for a given \((p, \mu)\), if the condition in part (i) holds \((D(p, \mu) \leq x_{ID})\) then patients do not prefer 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.

### 3.1.2. Objective function and optimal decisions

Let \( \lambda_t(p, \mu) \) denote the equilibrium arrival rate of telemedicine visits and \( \lambda_i(p, \mu) \) for in-person visits per unit time, as defined in (26) and (27), for a given price \( p \) and a service rate \( \mu \). Most of the cost incurred by a specialist will be fixed in nature (cost of the facility, staff and so on). Most specialists devote a certain proportion of their time for 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 \( \mu \), then the revenue function \( \bar{R} : \mathbb{R}^2_+ \rightarrow \mathbb{R} \) for the specialist is defined by

\[
\bar{R}(p, \mu) = \begin{cases} 
p(\lambda_t(p, \mu) + \lambda_i(p, \mu)), & \text{if } p \geq 0, \\
0, & \text{otherwise} \end{cases}
\]  

(32)

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

\[
\bar{R}^* = \sup_{p \geq 0, \mu > 0} \bar{R}(p, \mu).
\]  

(33)

Let \((p^*, \mu^*)\) denote the optimal decisions for the specialist (their existence will be proven below).

**Separability:** We next show that when the specialist does not differentiate between the 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 optimal decisions of the specialist. The separability results from the fact that the difference between the utilities, as given by (28), 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 a specialist offers only the in-person mode or only the telemedicine mode then his (revenue-maximizing) optimal actions are determined using Lemma 1 and Theorem 1. Let \((p_k^*, \mu_k^*)\) denote the optimal decisions for a specialist offering only in-person mode for \( k = 1 \) and only the telemedicine mode for \( k = 2 \). We similarly use subscript
\(\lambda_k(p, \mu) = \Lambda F(D(p, \mu))\) and \(\lambda_k(p, \mu) = \Lambda F(x_{TM}(p, \mu))\).

\[(34)\]

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.

**Theorem 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), \(p^* = p^*_m\) and \(\mu^* = \mu^*_m\). Hence the total arrival rate satisfies \(\lambda_i(p^*, \mu^*) + \lambda_t(p^*, \mu^*) = \lambda^*_m\).

From Theorem 4, 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 actions as the one that obtains more revenue.

We can also identify the ensuing equilibrium if the specialist chooses the optimal actions given in Theorem 4 using Lemma 9 in the appendix. Specifically, if \(R^*_1 \geq R^*_2\) and the specialist chooses the parameters \(p^* = p^*_1\) and \(\mu^* = \mu^*_1\), 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 3(ii).

### 3.1.3. Feasibility of telemedicine

In this section we explore the effect of offering the telemedicine mode on the specialist’s optimal actions. We have the following proposition.

**Proposition 3.** Assume that \(\lambda^*_i > 0\). If

\[s < \alpha (m_t - m_i + t \left(F^{-1}(\lambda^*_i/\Lambda)\right))\],

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, \(\lambda^*_i + \lambda^*_t > \lambda^*_i\); 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, \(\mu^* > \mu^*_i\).

The proposition gives a simple necessary condition to check the feasibility of telemedicine. 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 \(m_i\), the treatment benefit from telemedicine. By part (ii) of Proposition 3, the specialist sees more patients per unit time and
also at a faster rate. Thus, the results from our model support the hypothesis that telemedicine increases patients’ access to specialists; see [Kvedar et al. (2014)]. Interestingly, the increased access comes from increasing the “efficiency” in the system by means of an increased service rate. Even though the specialist spends less time with each patient on average, the patient’s satisfaction or net utility increases because of the reduced travel burden. Recent work by [Dorsey et al. (2013)] also offers empirical support that telemedicine could increase specialist efficiency through shortened visit times. However, it is to be noted that patients who continue to choose the in-person mode of treatment will experience a reduced net utility because of the faster service rate.

3.1.4. 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 Propositions 2 and 3 to make a conclusion because of the differences in the current model compared with that studied in Section 2.4.

First we define the total welfare when some patients choose telemedicine. Due to the independence of $x_{ID}$ from $\mu$ and $p$ and Lemma 6, if the specialist chooses arrival rate $\lambda$, then those patients that are located at a distance between 0 and $(x_{ID} \land F^{-1}(\lambda/\Lambda))$ receive higher utility from the in-person mode, and those located between $(x_{ID} \land F^{-1}(\lambda/\Lambda))$ and $F^{-1}(\lambda/\Lambda)$ from the telemedicine mode. Therefore, the total welfare, if the specialist chooses to serve an arrival rate of $\lambda$ and uses a service rate of $\mu$, is given by

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

(36)

where

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

(37)

and $\tilde{m} = \alpha m_i + (1-\alpha)m_i$.

Our goal is to show that if telemedicine is feasible, and the specialist chooses the revenue-maximizing arrival and service rates, then the total welfare is greater than the total welfare if the specialist offers only the in-person mode and chooses the revenue-maximizing arrival and service rates. That is,

$$U_d(\lambda^*_2, \mu^*_2) \geq U_1(\lambda^*_1, \mu^*_1),$$

(38)

where $\lambda^*_i$ and $\mu^*_i$ are defined as in the previous section for $i = 1, 2$ and $U_i(\lambda, \mu)$ denote the total welfare when the arrival rate is $\lambda$ and the service rate is $\mu$ for a specialist offering only in-person treatment defined as in (12). Note that when telemedicine is feasible by Proposition 3 the optimal arrival and service rates coincide with those when the specialist offers only telemedicine mode.
Proposition 4. If (35) holds, then (38) holds, and so the total welfare increases with the introduction of telemedicine.

The proof of Proposition 4 is still valid if one assumes $\lambda^*_s \geq \Lambda F(x_{ID})$ instead of (35), where $\lambda^*_s$ is the socially optimal arrival rate given in Theorem 2 for a specialist offering only the in-person mode.

3.2. Treatment model using distinct service rates and prices

Next we consider a different model for the specialist’s approach to the telemedicine mode based on our observations in practice, where the specialist allocates a certain 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 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 service modes together.

In this case it does not seem to be possible to obtain a result corresponding to Theorem 4, so a closed-form solution for the optimal actions of the specialist is not available. Hence we make the following additional 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 $\mu$ is the service rate and $\delta > 0$ represents the proportionality constant.

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

The specialist has five variables to optimize: service rates and prices for each mode, and the amount of time dedicated to each mode. Because total enumeration is inefficient in this case due to the size of the problem, we developed a method explained in [Rajan et al. (2014)] to identify the optimal actions of a specialist given $r$. Then we used a numerical search algorithm to find the optimal $r$. We next summarize our findings from numerical experiments.

As a base case, we considered the following values for the various parameters in our model: $\delta = 1; \Lambda = 1; m_i = 60; m_t = 60; t_o = 10; \beta = 0.1; s = 10; \alpha = 0.75; c = 5$. We also carried out a sensitivity analysis by allowing $m_i$ and $\frac{t_o}{\Lambda}$ to vary from 0 to 90 and 0 to 30 with 10 equal increments, keeping other values as in the base case. This gave us 100 combinations. Similarly, we considered 10 different values for $\alpha$ from 0.01 to 1 with equal increments and for $s$ from 0 to 45 with equal increments.
Figures 1a and 1b compare the Revenue for the specialist before Telemedicine (Rev B) and after Telemedicine (Rev TM). We observe that as the feasibility of telemedicine ($\alpha$) or the benefit from a telemedicine visit ($m_t$) increases when compared to the benefit from an in-person visit ($m_i$), the proportion of time spent for 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. Region I is when telemedicine is not offered by the specialist and Region II is when the specialist offers telemedicine mode.

Figures 1c and 1d compare the Service rates for the specialist before Telemedicine ($\mu$ BTM) and after Telemedicine ($\mu$ ATM and $\mu$ t ATM). We can see that the overall service rate increases with telemedicine. However, we also find that the optimal service rate for telemedicine can be lower (Region II) or higher (Region III) than the optimal service rate for in-person visits. The service rates for telemedicine visits are lower than in-person service rates only at lower values for the benefit from telemedicine visits ($m_t \leq 40$) or the feasibility factor for telemedicine ($\alpha \leq 0.6$).

Figures 1e and 1f compare the prices for the specialist before Telemedicine ($p$ BTM) and after Telemedicine ($p$ i ATM and $p$ t ATM). We see that both in-person and telemedicine prices in general drop after the introduction of telemedicine. Only if the benefit from telemedicine visits is significantly larger than the benefit from in-person visits ($m_t >> m_i$, Region III Figure 1f), it is optimal for the specialist to charge higher prices for telemedicine visits.

Figures 1g and 1h compare the utilization levels for the specialist before Telemedicine (Util BTM) and after Telemedicine (Utili ATM and Utilt ATM). We can see that the utilization levels go up, with more patients getting treatment than before. Hence the congestion cost or waiting cost for all patients increases when the specialist can choose two different service rates for in-person and telemedicine visits.

3.2.1. **Effect on total welfare** Figures 2a and 2b compare the Total Welfare for a Revenue-Maximizing specialist before Telemedicine (TWRM), Total Welfare for a Welfare-Maximizing specialist before Telemedicine (TWWM) and Total Welfare for a Revenue-Maximizing specialist after Telemedicine (TMTWRM). Comparing TWRM and TWWM, we can see that there is no efficiency loss when customers are homogeneous ($t = 0$). The efficiency loss increases with degree of heterogeneity as seen in Figure 2a. The efficiency loss is also fully adjusted for, if the transportation cost is too high (Region III Figure 2a).

We also observe that, even for a revenue-maximizing specialist, the total patient welfare increases when the specialist offers telemedicine treatment. Thus, even though congestion costs increase for the patients who were undergoing treatment before telemedicine was made available, reducing their
(a) Revenue and $r$ vs $\alpha$: $m_1 = 60; t = 10$

(b) Revenue and $r$ vs $m_2$: $\alpha = 0.75; t = 10$

(c) Service rates vs $\alpha$: $m_1 = 60; t = 10$

(d) Service rates vs $m_2$: $\alpha = 0.75; t = 10$

(e) Prices vs $\alpha$: $m_1 = 60; t = 10$

(f) Prices vs $m_2$: $\alpha = 0.75; t = 10$

(g) Utilization vs $\alpha$: $m_1 = 60; t = 4$

(h) Utilization vs $m_2$: $\alpha = 0.75; t = 4$

Figure 1  Numerical Analysis: $\delta = 1; m_i = 60; \beta = 0.1; s = 10; c = 5$
welfare, the increase in welfare because new patients seek treatment more than compensates for this loss.

It is also interesting to note that the total welfare under a revenue-maximizing specialist who introduces telemedicine is more than the welfare under a welfare-maximizing specialist who does not offer telemedicine if the transportation cost is high (Region III Figure 2a) or if the benefit from telemedicine visit is not too low (Region III Figure 2b). In fact even if the benefit from telemedicine visits is less than that of in-person visit \( m_t < m_i \), the total welfare after introducing telemedicine can be more than that of the case without telemedicine owing to the reduced transportation costs (Region III Figure 2b). When the transportation cost or benefit from telemedicine visit is in Region II, the total welfare under telemedicine is more when you consider the revenue-maximizing specialist but less than that under a welfare maximizing specialist.

![Figure 2 Welfare Analysis](image)

\[(a) \text{ Welfare vs } t: \delta = 1; m_i = 60; m_t = 39; \beta = 0.1; s = 10; \alpha = 0.75; c = 5\]

\[(b) \text{ Welfare vs } m_t: \delta = 1; m_i = 60; t = 10; \beta = 0.1; s = 10; \alpha = 0.75; c = 5\]

4. Conclusions and Policy Implications

In this paper, we consider an operating model of a specialist treating chronically ill patients arriving from different locations. Our research extends the analytical results of service interactions for 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 price and service rate decisions by the specialist where the heterogeneity could be due to different health conditions or the travel burden each patient faces. We use these new results in analyzing the economic and operational implications of using telemedicine to manage patients suffering from chronic conditions. The model especially allows us to explore the impact of introducing telemedicine on
service quality, specialist productivity and income, population coverage, cost of care and patient welfare.

We obtain the following three important managerial insights in the case before telemedicine is introduced in the model. First, in order to maintain quality of service, it is optimal for both kinds of specialists (revenue-maximizing and social welfare-maximizing) to accelerate the service rate at a slower pace relative to the increase in the expected patient workload. Our second managerial insight addresses the case of a specialist who is assigned an exogenous work load. In this case, both the revenue-maximizing and the welfare-maximizing specialists find it optimal to spend, on average, the same amount of time with each patient.

Third, the arrival rate in a given period for a revenue-maximizing specialist is always lower than (or equal to) the number seen by a welfare-maximizing specialist. On the other hand, the former spends more time with each patient than the latter. As a result, the congestion costs for the revenue-maximizing specialist are always lower than they are for the welfare-maximizing specialist.

With the introduction of telemedicine technology patients’ strategic choice falls into one of four mutually exclusive outcomes: 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 exclusively via telemedicine, and patients (for instance, those who live even farther away) who were not treated before choose to stay untreated by that specialist. We find that with the introduction of telemedicine, social welfare increases even for the population served by a specialist who maximizes his own revenue.

Our research highlights some very interesting policy implications with respect to the deployment of and reimbursement for using telemedicine technology for treating chronically ill patients. While the clinical efficacy of using telemedicine in this context has been proven in prior studies by us and others, there still remains a host of issues that prevent it from being widely accepted.

Our analytical results clearly show that though patients might incur an additional cost for technological support, chronically ill patients are going to benefit from the introduction of telemedicine in terms of a reduced travel burden and a lower fee for service. Some patients who have not been treated before will gain access to specialists as telemedicine increases the geographical coverage of providers. However, the benefit from telemedicine is not uniform for all patients, and indeed some patients, unexpectedly, will be even worse off with the introduction of telemedicine. Despite this fact, total welfare will increase even when specialists remain revenue-maximizers.

Our results also indicate that telemedicine benefits the specialist physicians as they enjoy higher productivity and higher revenue. Currently, one barrier for the deployment of telemedicine is the
lack of clear guidelines for specialist reimbursement. Assuming that 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) will always be lower than the fees levied before telemedicine, thereby benefiting both patients and third-party payers. Yet the specialists are expected to be financially better off too, since they will earn higher revenue with the introduction of telemedicine.

Our analysis leads to the following policy implications to remove some of the administrative barriers for the implementation of telemedicine, see Weinstein et al. (2014). One major issue is the issue of telemedicine’s efficacy. It is important to recognize that not all specialist groups or patient populations will benefit equally. The degree of cross-sectional heterogeneity among patients in terms of medical condition or travel burden has a significant impact on the marginal benefit from the introduction of telemedicine. This implies, for instance, that telemedicine provides greater benefits for both patients and physicians when employed to treat a highly heterogeneous patient population, such as patients in rural areas or those with serious motion disorders both of which cause relatively higher travel burdens.

Another barrier to the widespread adoption of telemedicine is the cost of the necessary technology. Our results indicate that it is optimal for the specialist physicians to indirectly subsidize part of these costs, and for the patients also to absorb a portion of these technology expenses. Another deployment barrier has to do with reimbursement (see Zanaboni and Wootton (2012)). Our results suggest that specialist physicians and patients will both benefit from telemedicine - even without any subsidies - so long as the specialists are reimbursed for telemedicine visits, even if the new (uniform) fees per face-to-face or remote visit are lower than the current fees for face-to-face office visits. In fact, several state Medicaid programs have already moved forward allowing for proper reimbursement of telemedicine services (see Thomas and Capistrant (2014)). For example, the New York Governor signed a legislation that would allow some licensed health providers in New York to get reimbursed for telemedicine visits effective January 1, 2016.

References


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


Shone, R., V.A. Knight, J.E. Williams. 2013. Comparisons between observable and unobservable M /M /1 queues with respect to optimal customer behavior. *European Journal of Operational Research* 227(1) 133–141.


Appendix

A. List of Notations

- $x$: Patient index, Distance of the patient from the specialist
- $X_m$: Upper bound on patient index, the maximum distance from the specialist
- $m$: Reward per visit to the specialist, subscripts “i” and “t” 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
- $\Lambda$: Total arrival rate if the specialist decides to serve all patients
- $p$: Price charged by a revenue maximizing specialist
- $\mu$: Average service rate of revenue maximizing specialist
- $\lambda$: Arrival rate of patient visits to the specialist
- $\beta$: Co-insurance rate (a fraction between 0 and 1)
- $Q(\mu)$: The cost or disutility to the patient as a function of the service rate
- $E_W$: The expected time spent by the patient in the system
- $c$: Opportunity cost per unit time for the patients
- $A = \{i, t, o\}$: Action set for the patients “i” denotes (in-person) treatment, “t” denotes telemmedicine treatment and “o” denotes no treatment
- $g_x$: Distribution on $A$ for a patient indexed by $x$
- $t(x)$: Transportation cost (perceived) for the patient located at distance $x$ from specialist
- $R(p, \mu)$: Revenue function for the revenue-maximizing specialist who does not offer telemedicine defined in (7)
- $U(\lambda, \mu)$: Total utility function for the welfare-maximizing specialist defined in (12)
- $s$: Amortized cost to the patient for setting up telemedicine visits
- $\alpha$: The proportion of visits possible via telemedicine. Thus for $(1 - \alpha)$ fraction of the total visits, the patient has to visit the hospital in-person (including emergency visits)
- $\lambda_{im}$: Arrival rate of in-person mode patients to the specialist
- $\lambda_{tm}$: Arrival rate of telemedicine mode patients to the specialist
- $\lambda_i$: Arrival rate of in-person visits to the specialist
- $\lambda_t$: Arrival rate of telemedicine visits to the specialist
- $D(p, \mu)$: The location of the patient who is indifferent between choosing treatment and not seeking treatment when the telemedicine mode is not available.
- $x_{ID}$: The location of the patient who is indifferent between the two modes of treatment.
- $x_{TM}$: The 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 (32)
B. Proofs of the results in Section 2

B.1. Proofs of the results in Section 2.1.4

Proof of Lemma 1: We have assumed that $F$ 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 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 $F'(x) > 0,$

$$H'(u) = \frac{1}{F'(H(u))}$$

and

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

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

$$H'(u) > 0 \text{ and } H''(u) \geq 0 \quad \forall u \in [0, 1]. \quad (39)$$

If the arrival rate is $\lambda$, then the location of the farthest patient who seeks treatment is given by $H(\lambda/\Lambda)$. Since the marginal utility of that patient from treatment is $V'\left(\lambda\right)$,

$$V'(\lambda) = m - t(H(\lambda/\Lambda)). \quad (40)$$

Hence, for all $\lambda \in [0, \Lambda]$,

$$V''(\lambda) = -\frac{1}{\Lambda} t'(H(\lambda/\Lambda)) H'(\lambda/\Lambda),$$

and

$$V'''(\lambda) = -\frac{1}{\Lambda^2} \left[ (H'(\lambda/\Lambda))^2 t''(H(\lambda/\Lambda)) + t'(H(\lambda/\Lambda)) H''(\lambda/\Lambda) \right].$$

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 (39) and the assumption $t'(x) > 0, V''(\lambda) \leq 0$; by the fact that $t''(x) \geq 0$, and by (39), $V'''(\lambda) \leq 0$ on $\lambda \in [0, \Lambda F^{-1}(M)]$. □

B.2. Proofs of the results in Section 2.2

Proof of Lemma 2: Let $\mu \geq Q^{-1}(M_v)$. We note that the inverse exists by Assumptions B.1, B.2, and (14). Then, $Q(\mu) \geq M_v$. Let $\mu$ be such that $\mu > \lambda$. If not, $\bar{R}(\lambda, \mu) = 0$ from (7). Since $c > 0$ and $\mu > \lambda$, we have $Q(\mu) + \frac{c}{\mu - \lambda} > M_v$, which implies, from (6), that $x^* = 0$, and so $p(\lambda, \mu) = 0$ from (15). Hence $\bar{R}(\lambda, \mu) = 0$ from (16). Now we find the lower limit for $\mu$. Let $\mu \leq \lambda + \frac{c}{M_v}$. Rearranging
terms, we have, and since \( Q(\mu) \geq 0, Q(\mu) + \frac{c}{\mu - \lambda} \geq M_v \), which implies, from (6), that \( x^* = 0 \), and so \( p(\lambda, \mu) = 0 \) from (15).

Finally, if \( \lambda \geq \kappa \), then

\[
m - t(F^{-1}(\lambda/\Lambda)) \leq Q(c/M_v). \tag{41}
\]

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 (41), gives \( m - t(F^{-1}(\lambda/\Lambda)) - Q(\mu) \leq 0 \). Thus \( p(\lambda, \mu) = 0 \) by (15), giving the desired result. \( \Box \)

**Proof of Lemma 2:** Let \( \mu' > 0 \) and \( p' > 0 \) be such that \( R(\mu', \mu') > 0 \) (if \( \mu' = 0 \) or \( p' = 0 \) then \( R(\mu', \mu') = 0 \) by (5), (6) and (7)). If \( R(\mu', \mu') > 0 \), then \( \lambda(\mu', \mu') > 0 \), and so \( x^* > 0 \) by (5). In addition \( x^* < X_{\mu'} \) by (6). From Assumptions A.1 and A.2, \( F^{-1}(t(.)) \) is continuous on \((0,1)\) also. This implies by (5) and (6) that

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

and \( \mu' > \lambda(\mu', \mu') \). Thus, by (15), \( p(\lambda(\mu', \mu'), \mu') = p' \). Hence \( \tilde{R}(\lambda(\mu', \mu'), \mu') = R(\mu', \mu') \).

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 (16)). This implies by (16) that \( p(\lambda', \mu') > 0 \) and \( \lambda' < \mu' \). Hence by Lemma 2 \( \lambda' < \Lambda \) and by (15)

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

By (6), \( \lambda(p(\lambda', \mu'), \mu') = \lambda' \). Hence \( \tilde{R}(\lambda', \mu') = R(\lambda(p(\lambda', \mu'), \mu')) \). The fact that \( \tilde{R}^* = R^* \) follows from (19) and (20). \( \Box \)

**Proof of Lemma 2:** Let \( M > 0 \) denote a finite constant. Define

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

By Assumptions B.1 and B.2 \( Q' \) is non-decreasing and continuous. Also because \( \frac{c}{(\mu - \lambda)^2} \) is a decreasing continuous function for \( \mu > \lambda \) with its values dense in \((0, \infty)\), and from Assumptions B.1 and B.2 there exists \( \mu_\lambda > \lambda \) such that \( J(\lambda, \mu_\lambda) = 0 \).

Because \( Q' \) is non-decreasing, \( \mu_{\lambda_1} < \mu_{\lambda_2} \), if \( \lambda_1 < \lambda_2 \). Therefore,

\[
\mu_\lambda - \lambda \geq \sqrt{c/Q'(\mu_0)}, \text{ for all } \lambda \in [0, M]. \tag{42}
\]

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

\[
J(\lambda, \gamma(\lambda)) = 0, \text{ for } \lambda \in (0, M). \tag{43}
\]
Before we proceed with the proof of (43) we recall the implicit function theorem (IFT). The IFT states the following. If \( J \) is defined on an open disk containing \((\lambda, \mu)\), where \( J(\lambda, \mu) = 0 \), \( J_\mu(\lambda, \mu) \neq 0 \), and \( J_\mu \) and \( J_\lambda \) are continuous on the disk, then the equation \( J(\lambda, \mu) = 0 \) defines \( \mu \) as a function of \( \lambda \); i.e., there exists a function \( \gamma \) such that \( J(\lambda', \gamma(\lambda')) = 0 \) in a neighborhood of \((\lambda, \mu)\). We now check whether all the conditions hold, 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 \( \lambda \) and \( \mu \), \( J_\lambda \) and \( J_\mu \) 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 [B.2] 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 \( \gamma_\lambda \) and a neighborhood \( A_\lambda \) of \( \lambda \) such that \( J(\lambda', \gamma_\lambda(\lambda')) = 0 \) for all \( \lambda' \in A_\lambda \).

Although the IFT proves the existence of \( \gamma_\lambda \) we still have not proved that there exists a unique differentiable function \( \gamma \) on \((0, M)\) and that \( \gamma(\lambda) > \lambda \). We next prove that for \( \lambda \neq \lambda' \), \( \gamma_\lambda = \gamma_{\lambda'} \) on \( A_\lambda \cap A_{\lambda'}(\neq \emptyset) \), proving uniqueness. The fact that \( \gamma(\lambda) > \lambda + \delta \) for some \( \delta > 0 \) follows from (42).

First, each \( \gamma_\lambda \) is continuous and differentiable by the IFT. Assume that there exists \( \lambda' \in A_\lambda \) such that \( \gamma_\lambda(\lambda') < \lambda' \) and \( \lambda' < \lambda \). This implies by the continuity of \( \gamma_\lambda \) 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 [B.1] and [B.2]. Hence no such \( \lambda' \) exists. We can prove the conclusion for \( \lambda' > \lambda \) similarly. Hence for all \( \lambda' \in A_\lambda \), \( \gamma_\lambda(\lambda') > \lambda' \). However, there is a unique \( \gamma_\lambda(\lambda') > \lambda' \) that satisfies (43), so \( \gamma_\lambda = \gamma_{\lambda'} \) on \( A_\lambda \cap A_{\lambda'}(\neq \emptyset) \). Therefore there is a unique \( \gamma \) that satisfies the conditions of the result.

By the IFT we have

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

From Assumption [B.2] and the fact that \( \gamma(\lambda) > \lambda + \delta \), \( \gamma'(\lambda) > 0 \) and \( \gamma'(\lambda) \leq 1 \) for all \( \lambda \in (0, M) \). \( \square \)

**Proof of Lemma** [2] Let \( \lambda \in (0, \kappa) \) and assume that there exists \( \mu > 0 \) such that \( \tilde{R}(\lambda, \mu) > 0 \). Next we show that given \( \lambda \), \( \gamma(\lambda) \) gives the optimal solution for (21). Let

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

and (44) and (45)
for \( \lambda \in [0, \Lambda] \) and \( \mu > \lambda \).

Note that

\[
p^+ (\lambda, \mu) \leq p(\lambda, \mu) \quad \text{and} \quad \tilde{R}^+ (\lambda, \mu) \leq \tilde{R}(\lambda, \mu).
\]

The first inequality in (46) follows from the definition of \( p(\lambda, \mu) \) (see (15)), and the second inequality follows from (46). We note that if \( \tilde{R}^+ (\lambda, \mu) < 0, \forall \mu > \lambda \), then \( p^+ (\lambda, \mu) < 0, \forall \mu > \lambda \) by (46) and thus \( p(\lambda, \mu) = 0, \forall \mu > \lambda \) from (15). Then \( \tilde{R}(\lambda, \mu) = 0, \forall \mu > \lambda \) from (16). 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

\[
\tilde{R}^+ (\lambda, \mu) = \tilde{R}(\lambda, \mu), \quad \text{if} \quad p^+ (\lambda, \mu) > 0 \text{ or } p(\lambda, \mu) > 0.
\]

(47)

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

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

We use KKT necessary conditions to obtain the optimal service rate. Given \( \lambda > 0 \), since the constraint \( \mu \geq \lambda + \frac{c}{M_v} \) is linear in \( \mu \), regularity conditions are satisfied, so if \( \mu^*_\lambda \) is optimal then there exists \( \sigma \) such that the following KKT necessary conditions are satisfied (see Proposition 6.5, Sundaram (2007)):

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

We then have the following two cases for non-negative \( \sigma \):

**Case 1, \( \sigma = 0 \)**: This implies that \( \mu^*_\lambda \) along with \( \sigma = 0 \) will satisfy the KKT conditions if \( \mu^*_\lambda \geq \lambda + \frac{c}{M_v} \).

Also, from (46), \( \frac{\partial \tilde{R}^+}{\partial \mu} = -J(\lambda, \mu) \), Hence \( \mu^*_\lambda = \gamma (\lambda) \) is the unique solution.

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

Next we argue that we can ignore the solution given by Case 2 for the optimization problem (21). Note that if \( \mu^*_\lambda = \left( \lambda + \frac{c}{M_v} \right) \) (i.e. Case 2 holds), from Lemmas 2 and 3, 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 \( \mu^*_\lambda \) as in Case 1, since \( \tilde{R}(\lambda, \mu) \geq 0 \). Otherwise, if \( \tilde{R}^+ (\lambda, \mu^*_\lambda) > 0 \), then \( \mu = \left( \lambda + \frac{c}{M_v} \right) \) cannot be optimal, so we can only consider Case 1. Therefore, by (47), giving 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 2. \( \square \)
Proof of Theorem 2: 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 (16) because \(\tilde{R}(\lambda, \mu) > 0\). If \(\tilde{R}(\lambda, \mu) > 0\) for some \(\lambda\) and \(\mu\) 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 (47) and Lemma 3, \((\lambda^*, \mu^*)\) must be the solution of (18) as well. We next show that such \(\lambda^*\) and \(\mu^*\) exists.

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 2 and the second inequality follows from (47) and Lemma 5. By (45) and Lemma 5, \(\tilde{R}^+(\lambda, \gamma(\lambda))\) is continuous on \((0, \Lambda]\) and \(\limsup_{\lambda \to 0} \tilde{R}^+(\lambda, \gamma(\lambda)) \leq 0\), hence there exists \(0 < \epsilon < \kappa\) such that \(\tilde{R}^+(x, \gamma(x)) < \tilde{R}^+(\lambda, \gamma(\lambda))\) for \(0 < x \leq \epsilon\). Hence

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

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 4, \(\tilde{R}^+(\lambda, \gamma(\lambda))\) is continuous in \(\lambda\) 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 \(\lambda\) on \([\epsilon, \kappa]\). First, by (43),

\[
\beta \frac{d \tilde{R}^+(\lambda, \gamma(\lambda))}{d\lambda} = V'(\lambda) - Q(\gamma(\lambda)) - \frac{c}{\gamma(\lambda) - \lambda} + \lambda V''(\lambda) - \frac{e}{\gamma(\lambda) - \lambda} \left( \frac{c}{(\gamma(\lambda) - \lambda)^2} \right)
\]

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 4, Lemma 5 and Assumptions A.1 A.2 B.1 and B.2 each of the terms is either negative or non-positive. Also, by Lemma 4,

\[
\beta \frac{d^2 \tilde{R}^+(\lambda, \gamma(\lambda))}{d\lambda^2} < 0, \text{ for } \lambda \in (0, \kappa).
\]
Thus, \( \tilde{R}^+(\lambda, \gamma(\lambda)) \) is concave in \( \lambda \), 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 \( \lambda \) 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 conditions for an optimal solution. The necessary first-order condition with \( \tilde{R}^+(\lambda, \gamma(\lambda)) \) gives \( \tilde{R}(\lambda) \). Uniqueness is then ensured by strict concavity. Also \( \tilde{R}(\lambda) \) has a solution, \( \tilde{\lambda} \in [\epsilon, \kappa] \), because the optimal point is not at the boundaries. Second part of the theorem is obvious. \( \square \)

**Proof of Proposition 1**: Let \( V_1 \) and \( V_2 \) be defined as in (11) when the travel burden is given by \( t_1 \) and \( t_2 \) respectively. Therefore, we have, \( V_1'(\lambda) = m - t_1(F^{-1}(\lambda/\Lambda)) \) and \( V_2'(\lambda) = m - t_2(F^{-1}(\lambda/\Lambda)) \).

Let \( h_1 \) be defined by

\[
h_1(\lambda) = m - Q(\gamma(\lambda)) - \frac{c}{\gamma(\lambda) - \lambda} - \frac{c\lambda}{(\gamma(\lambda) - \lambda)^2}.
\]

(52)

Also let \( h_2 \) be defined by,

\[
h_2(\lambda) = V_2'(\lambda) + \lambda V_2''(\lambda) - Q(\gamma(\lambda)) - \frac{c}{\gamma(\lambda) - \lambda} - \frac{c\lambda}{(\gamma(\lambda) - \lambda)^2}.
\]

(53)

From Assumption A.2 and (52)–(53) we have

\[
h_1(\lambda) \geq h_2(\lambda).
\]

(54)

By Theorem 1, \( h_2(\lambda^*) = 0 \), and by (51), \( h_2(\lambda) > 0 \) for \( \lambda \in (0, \lambda^*) \). Hence, \( h_1(\lambda) > 0 \) for \( \lambda \in (0, \lambda^*) \) by (54), giving part (i) by Theorem 2. The proof of \( \mu^*_s \geq \mu^* \) follows directly from part (i) and Lemmas 5 and 8. \( \square \)

**B.3. Proofs of the results in Section 2.3**

The following result corresponds to Lemma 2 in this setting.

**Lemma 7.** If \( \mu \geq Q^{-1}(M_v) \) or if \( \mu \leq \lambda + \frac{c}{M_v} \), then \( U(\lambda, \mu) \leq 0 \).

The proof is similar to that of Lemma 2 and thus is omitted. Using this result, the objective (13) can be written as

\[
U^* = \sup_{\lambda \geq 0, \frac{\lambda}{M_v} \leq \mu \leq Q^{-1}(M_v)} U(\lambda, \mu)
\]

and

\[
U(0, \mu) = 0
\]

(55)

for any \( \mu \geq 0 \) by (12).
Next, we find the optimal service rate, \( \mu \), given the arrival rate, \( \lambda \); that is, given \( \lambda \geq 0 \), we solve the following optimization problem:

\[
U^*(\lambda) = \sup_{\lambda^+, \mu^+ \leq \mu} U(\lambda, \mu).
\]

We then have the following lemma, which is similar to Lemma 5 for \( \gamma(\lambda) \) defined in (22).

**Lemma 8.** Given \( \lambda \in (0, \Lambda] \), if there exists \( \mu > \lambda \) such that \( U(\lambda, \mu) > 0 \), then

\[
U^*(\lambda) = U(\lambda, \gamma(\lambda)).
\] (56)

Next we prove Theorem 2.

**Proof of Theorem 2**. Assume that there exists \( (\lambda, \mu) \) such that \( U(\lambda, \mu) > 0 \). We have

\[
\sup_{0 \leq \lambda \leq \Lambda, \mu \geq 0} U(\lambda, \mu) = \sup_{0 < \lambda \leq \Lambda, \mu \geq 0} U(\lambda, \mu) = \sup_{0 < \lambda \leq Q^{-1}(M_v) \wedge \Lambda} U(\lambda, \gamma(\lambda)),
\] (57)

where the first equality follows from (12) and the second equality follows from the assumption that \( U^* > 0 \) and Lemma 8.

Define

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

Because \( U^*(\lambda^*) = \sup_{\lambda \geq 0} U^*(\lambda) \), and \( U^* > 0 \) by the assumption of the theorem

\[
\sup_{0 < \lambda \leq Q^{-1}(M_v) \wedge \Lambda} U(\lambda, \gamma(\lambda)) = \sup_{0 < \lambda \leq Q^{-1}(M_v) \wedge \Lambda} U^+(\lambda).
\]

Note that \( U^+ \) is continuous on \([0, \Lambda]\) and \( \lim_{\lambda \to 0} U^+(\lambda) \leq 0 \). Therefore similar to the proof of Theorem 1, there exists \( Q^{-1}(M_v) \wedge \Lambda > \epsilon > 0 \) such that

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

By Lemma 5 and Assumptions A.1, A.2, B.1, and B.2, \( U^+ \) is continuous. Therefore there exists \( \lambda^*_s \in [\epsilon, Q^{-1}(M_v) \wedge \Lambda] \) such that

\[
U^+(\lambda^*_s) = \sup_{\epsilon \leq \lambda \leq Q^{-1}(M_v) \wedge \Lambda} U^+(\lambda)
\]

and by (57) and (58), \( U^*(\lambda^*_s) = \sup_{\lambda \geq 0} U^*(\lambda) \).

Using the second derivative of \( U^+(\lambda) \), we next show that \( U^+(\lambda) \) is concave in \( \lambda \) on \((0, \Lambda]\). First, by (43),

\[
\frac{dU^+(\lambda)}{d\lambda} = V'(\lambda) - Q(\gamma(\lambda)) - \frac{c}{\gamma(\lambda) - \lambda} - \lambda \left( \frac{c}{(\gamma(\lambda) - \lambda)^2} \right).
\]
Also, again by (43),
\[
\frac{d^2U^+(\lambda, \gamma(\lambda))}{d\lambda^2} = V''(\lambda) - \frac{2c}{(\gamma(\lambda) - \lambda)^2} - \frac{2c(1 - \gamma'(\lambda))\lambda}{(\gamma(\lambda) - \lambda)^3}.
\]
Hence, since \(\gamma(\lambda) > \lambda\) by Lemma 4 and by Lemma 5 and Assumption A.2 each term is non-positive,
\[
\frac{d^2U^+(\lambda, \gamma(\lambda))}{d\lambda^2} < 0, \quad \forall \lambda \in (0, \Lambda]. \tag{58}
\]
Thus, \(U^+(\lambda)\) is strictly concave in \(\lambda\), for \(\lambda \in (0, \Lambda]\). So we can use the necessary first-order conditions for an optimal solution. Then if \(\tilde{\lambda}_x \in [0, \Lambda]\) exists, it must be the optimal solution due to strict concavity. Otherwise, the optimal solution must be at the boundary \(\Lambda\). The second part of the result follows from (55). □

C. Proofs of the results in Section 3

Proof of Theorem 3: First, by Lemma 6 and (28),
\[
\Psi_i(\lambda, p, \mu, x) > \Psi_t(\lambda, p, \mu, x), \quad \text{for all } x < x_{ID}, \tag{59}
\]
for any \(\lambda, p\) and \(\mu\). Hence in any equilibrium
\[
g_x(t) = 0 \text{ for } 0 \leq x < x_{ID}. \tag{60}
\]
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 \(\lambda_G\) denote the associated total arrival rate defined as in (27). Note that
\[
\lambda_G = \Lambda F(D(p, \mu)) \text{ by (27).}
\]
We first show that \(G\) is an equilibrium. By Lemma 6, (30), and (59), for \(x \in [0, D(p, \mu)]\), (2) holds with \(g_x(i) = 1\). By (29), (30), and (60), (2) holds with \(g_x(o) = 1\) for \(x \in (D(p, \mu), x_{ID}]\). Finally, by Lemma 6 and (30), (2) 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 \(\lambda'\) defined as in (27). First assume that \(g_x'(t) > 0\) for some \(x'\), so \(\Psi_i(\lambda, p, \mu, x') \geq 0\) by (2). Then, by (60), \(x' \geq x_{ID}\). Also, by \(\Psi_t(\lambda, p, \mu, x') \geq 0\), (25), and the fact that \(t\) is strictly increasing,
\[
\Psi_i(\lambda', p, \mu, x) > 0 \text{ for all } x < x'. \tag{61}
\]
and by (59) this implies that
\[
\Psi_t(\lambda', p, \mu, x) > 0 \text{ for all } x < x_{ID}. \tag{62}
\]
If \( x' = x_{ID} \), then \( G \) and \( G' \) must be equal a.e. by (60) and (62). Thus assume that \( x' > x_{ID} \). Then, again by (61) and (62), \( g'_x(i) = 1 \) for all \( x < x_{ID} \), and \( g'_x(t) = 1 \) for \( x \in (x_{ID}, x') \). Therefore \( \lambda' > \lambda \) by (26) and the fact that \( x' > x_{ID} \). So by (24) there exists \( x'' < D(p, \mu) \) such that

\[
\Psi_t(\lambda', p, \mu, x) < 0 \quad \text{for all} \quad x \in [x'', D(p, \mu)].
\]

(63)

Since \( D(p, \mu) \leq x_{ID} \) and \( x' > x_{ID} \), (63) contradicts (62); therefore no such \( x' \) exists and patients only seek in-person visits. It then readily follows from Lemma 6 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 \( x \in (x_{ID}, D(p, \mu)) \), we have

\[
\Psi_t(\Lambda F(D(p, \mu)), p, \mu, x) \geq \Psi_t(\Lambda F(D(p, \mu)), p, \mu, x) > 0,
\]

where the first inequality follows from the definition of \( x_{ID} \) and the second follows from the definition of \( D(p, \mu) \). By (64), (31) 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 \( \lambda_G \) denote the associated total arrival rate defined as in (27). 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 \( x_{TM}(p, \mu) > x_{ID} \), \( \Psi_t(\lambda_G, p, \mu, x) \geq 0 \) for all \( x \in [0, x_{ID}] \). Hence, for \( x \in [0, x_{ID}] \), (2) holds with \( g_x(i) = 1 \). Similarly, for \( x \in (x_{ID}, x_{TM}(p, \mu)) \), by (29) and (31), \( \Psi_t(\lambda_G, p, \mu, x) \geq \Psi_t(\lambda_G, p, \mu, x) \) and \( \Psi_t(\lambda_G, p, \mu, x) \geq 0 \). Hence, for \( x \in (x_{ID}, x_{TM}(p, \mu)) \), (2) holds with \( g_x(t) = 1 \). Finally, by (31), the fact that \( x_{TM}(p, \mu) > x_{ID} \), and Lemma 6 \( \Psi_t(\lambda_G, p, \mu, x) \leq \Psi_t(\lambda_G, p, \mu, x) < 0 \) for all \( x > x_{TM}(p, \mu) \). Thus, for \( x > x_{TM}(p, \mu) \), (2) 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 \( \lambda' \) defined as in (27). First we show that for \( y > x_{TM}(p, \mu) \), in any equilibrium \( g'_x(o) = 1 \). If not, then by (2) \( \Psi_t(\lambda', p, \mu, y) \geq 0 \), and so by Lemma 6

\[
\Psi_t(\lambda', p, \mu, x) > 0, \quad \text{for all} \quad x < y.
\]

(65)

This implies that \( g'_x(t) = 1 \) for all \( x \in (x_{ID}, y) \), so \( \lambda' > \lambda_G \). But then, by (31) and the fact that \( x_{TM}(p, \mu) > x_{ID} \), \( \Psi_t(\lambda', p, \mu, x) < \Psi_t(\lambda', p, \mu, x) < 0 \), for \( x \in (x_{TM}, y) \). This obviously contradicts (65), so no such \( y \) exists. This also implies that \( \lambda \leq \lambda_G \). Therefore by (31), (60), and Lemma 6 any equilibrium strategy \( g'_x \) must satisfy \( g'_x(i) = 1 \) for all \( x \in [0, x_{ID}] \), since \( \Psi_t(\lambda', p, \mu, x) > \Psi_t(\lambda', p, \mu, x) > 0 \), 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. \( \square \)
Proof of Theorem 4: We prove the result by showing that for any \( p \geq 0 \) and \( \mu > 0 \),
\[
\bar{R}(p, \mu) = \max \{ R_1(p, \mu), R_2(p, \mu) \}. \tag{66}
\]
Fix \( p \geq 0 \) and \( \mu > 0 \). Note that if \( R_i(p, \mu) \geq \lambda_i'(p, \mu) \) for \( i \in \{1, 2\} \) and \( i' = \{1, 2\} \setminus \{i\} \), we prove the following result below.

**Lemma 9.** 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 3(i), (34), and Lemma 9, \( \bar{R}(p, \mu) = R_1(p, \mu) \). If on the other hand \( R_1(p, \mu) < R_2(p, \mu) \), then by Theorem 3(ii), (34), and Lemma 9, \( \bar{R}(p, \mu) = R_2(p, \mu) \), proving (66). We complete the proof by proving Lemma 9.

Assume that
\[
\lambda_1(p, \mu) \geq \lambda_2(p, \mu). \tag{67}
\]
This implies by (34) that
\[
D(p, \mu) \geq x_{TM}(p, \mu). \tag{68}
\]
Therefore
\[
\Psi_i(\lambda_1(p, \mu), p, \mu, D(p, \mu)) \geq \Psi_i(\lambda_2(p, \mu), p, \mu, D(p, \mu)) \geq \Psi_i(\lambda_1(p, \mu), p, \mu, D(p, \mu)), \tag{69}
\]
where the first inequality follows from (30), (31), and (68), and the second inequality follows from (67). By Lemma 6 and (69), \( \Psi_i(\lambda_1(p, \mu), p, \mu, x) \geq \Psi_i(\lambda_1(p, \mu), p, \mu, x) \), for all \( x \leq D(p, \mu) \). Thus \( D(p, \mu) \leq x_{ID} \) by (29), proving the first part of the lemma.

Now assume that
\[
\lambda_1(p, \mu) < \lambda_2(p, \mu). \tag{70}
\]
This implies by (34) that \( D(p, \mu) < x_{TM}(p, \mu) \). Therefore
\[
\Psi_i(\lambda_1(p, \mu), p, \mu, D(p, \mu)) < \Psi_i(\lambda_2(p, \mu), p, \mu, D(p, \mu)) \leq \Psi_i(\lambda_1(p, \mu), p, \mu, D(p, \mu)), \tag{71}
\]
where the first inequality follows from (30), (31), (70), and Lemma 6 and the second inequality follows from (70). By Lemma 6 and (71), \( \Psi_i(\lambda_1(p, \mu), p, \mu, x) < \Psi_i(\lambda_1(p, \mu), p, \mu, x), \) for all \( x \geq D(p, \mu) \). Thus \( D(p, \mu) > x_{ID} \) by (29), proving the second part of the lemma and concluding the proof. \( \square \)
Proof of Proposition 3: The proof follows from Theorems 3 and 4. Assume that $\lambda^*_1 > 0$ and (35) holds. By (34), $\lambda^*_1 = \Lambda F(D(p^*_1, \mu^*_1))$. Then by (28), (30), and (31),

$$\Psi_i(\lambda^*_1, p^*_1, \mu^*_1, D(p^*_1, \mu^*_1)) > \Psi_i(\lambda^*_1, p^*_1, \mu^*_1, D(p^*_1, \mu^*_1)) = 0.$$ 

Thus $x_{TM}(p^*_1, \mu^*_1) > D(p^*_1, \mu^*_1)$. Therefore, by (34), $R^*_2 > R^*_1$. Then, by Theorem 4, $R^*_2 = R^*_2$ and $\lambda^* = \lambda^*_2$, proving part (i). Part (ii) follows from Lemma 5. □

Proof of Proposition 4: If $\lambda^*_1 = 0$, then $U_1(\lambda^*_1, \mu^*_1) = 0$ by (12) as well, so (38) holds trivially. Assume that $\lambda^*_1 > 0$ and that (35) holds. If $U_d(\lambda, \mu) = 0$ for all $\lambda > 0$ and $\mu > 0$, then, by (37) and (11), $U_1(\lambda, \mu) = 0$ 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 8 it can be shown that if there exists $\mu > \lambda$ such that $U(\lambda, \mu) > 0$, then $U^*_d(\lambda) = U_d(\lambda, \gamma(\lambda))$.

If the specialist chooses to serve arrival rate $\lambda$, then the distance of the farthest patient is given by $x = F^{-1}(\lambda/\Lambda)$. By the proof of Proposition 3 if (35) holds, then $\lambda_1(\mu^*_1, p^*_1) \geq \Lambda F(x_{ID})$. Therefore, by Proposition 2, Theorem 2, and (58),

$$U^*_1(\Lambda F(x_{ID})) \geq U^*_1(\Lambda F(x)) \text{ for } x \leq x_{ID}. \quad (72)$$

Because $U^*_d(\Lambda F(x)) = U^*_1(\Lambda F(x))$ for $x \leq x_{ID}$, we have by (72) that

$$U^*_d(\Lambda F(x_{ID})) \geq U^*_1(\Lambda F(x)) \text{ for } x \leq x_{ID}. \quad (73)$$

Also, by (37),

$$\frac{dV^*_d(\lambda)}{d\lambda} = \bar{m} - (1 - \alpha)t(F^{-1}(\lambda/\Lambda)), \text{ for } \lambda > \Lambda F(x_{ID}).$$

Therefore, by (40),

$$\frac{dU^*_d(\lambda)}{d\lambda} = \frac{dU^*_2(\lambda)}{d\lambda}, \text{ for } \lambda > \Lambda F(x_{ID}). \quad (74)$$

By Proposition 2, Theorem 2 and (58),

$$\frac{dU^*_2(\lambda)}{d\lambda} > 0 \text{ for } \lambda < \lambda^*_2.$$ 

Thus, by (74), $U^*_d$ is increasing for $\lambda \in (\Lambda F(x_{ID}), \lambda^*_2)$. Combined with (73), this gives the desired result. □