Submitted to
manuscript (Please, provide the mansucript number!)

Improving Patient Access to Care: Performance Incentives and Competition in Healthcare Markets

Houyuan Jiang
Judge Business School, University of Cambridge, Cambridge, CB2 1AG, United Kingdom, h.jiang@jbs.cam.ac.uk

Zhan Pang
College of Business, City University of Hong Kong, Hong Kong, zhan.pang@cityu.edu.hk

Sergei Savin
The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, savin@wharton.upenn.edu

Performance-based compensation is gaining popularity as a mechanism for incentivizing providers of healthcare services to improve the quality of patient care. This paper investigates the effects of introducing performance-based incentives in a competitive healthcare market. In particular, we consider a market in which a payer (e.g. a government agency) applies a compensation contract to competing healthcare service providers in order to achieve a certain level of patient access to care, as measured by the expected time patients have to wait to receive care. In our model, we use $M/M/1$ queueing dynamics to describe patient service processes and assume that patient demand for care delivered by a particular provider is increasing in the level of access to care the provider ensures and decreasing in the levels of access to care at competing providers. Our analysis indicates that the presence of competition between providers may significantly alter the intended effect of performance-based incentives. In particular, we show that the joint effect of incentives and competition depends on two factors: 1) the aggressiveness of patient access targets that the payer imposes on providers, and 2) patient sensitivity to the level of access to care.

When the payer uses a "soft" approach to performance-based compensation by incentivizing but not requiring that providers reach an access-level target, the incentives and competition can produce opposing effects on patient access to care when aggressive service-level targets are used in the presence of access-sensitive patients or when moderate service-level targets are introduced in environments where patients a exhibit low degree of sensitivity to the level of access to care. In particular, we show that while moderate service-level targets can lead to an improvement in patient access to care when applied to a monopolistic provider, competition in settings with access-insensitive patients may diminish or even reverse this improvement. Under the "strict" approach to performance-based compensation, when the payer designs performance incentives to minimize the cost of imposing a common access-level target on all providers, the impact of competition on the level of incentivization required is also influenced by the patient population type: for access-sensitive patients, competitive pressure lowers the level of incentivization required to achieve a particular level of patient access to care, while for patients with low access sensitivity the effect of competition is to increase the incentivization level required. At the same time, the reduction in payers' costs resulting from the presence of competition is more pronounced in environments with access-insensitive patients.

Key words: Healthcare competition; waiting time target; performance-based incentives

1. Introduction and Literature Review

Faced with increasing pressure to contain rising costs while maintaining high-quality care, healthcare systems in the US and other developed countries are experimenting with policies that incentivize care providers to compete for patients and tie provider compensation to the quality of care delivered. In the US, the Patient Protection and Affordable Care Act (also known as the Affordable Care Act or ACA) introduced in 2010 (PPACA 2010) has created a way for tens of millions of new patients to access care and, at the same time, introduced a number of new approaches designed to slow down the rise in the overall cost of care.

It has long been argued that both the "prices" charged for healthcare services in the US and the"volume" of services delivered per capita could be the main culprits behind overall healthcare costs as well as their growth.

On the "price" side, while growth in health expenditure per capita in the US has slowed from an annual average of 2.3% over 2005–2009 to 1.5% over 2009–2013, health expenditure of more than $\$8,700$ per capita in the US continues to be the highest among the Organization for Economic Cooperation and Development (OECD) countries and is about 40% higher than that of closest rival Switzerland and 2.5 times higher than the OECD average, adjusted for purchasing power (OECD 2015). In one of most influential articles on the subject, Anderson et al. (2003) argue that the concentration of bargaining power on the provider side of the market may be one of leading reasons for this phenomenon: "Although the huge federal Medicare program and the federal-state Medicaid programs do possess some monopsonistic purchasing power, and large private insurers may enjoy some degree of monopsony power as well in some localities, the highly fragmented buy side of the US health system is relatively weak by international standards."

On the "volume" side, the research conducted by the Dartmouth Atlas Project (DAP) over the last two decades has identified persistent geographical variability in the utilization of healthcare service capacity for treating the same medical conditions (for some of the latest reports, see Wennberg et al. 2006 and Bynum et al. 2016). John Wennberg, the founder and driving force of the DAP, and his colleagues argue that much of this variability is "unwarranted," stemming from the underutilization of evidence-based practices, shortage of informed patient choices, and proliferation of "supply-driven" diagnostic and treatment decisions (see, for example, Wennberg 2002). In particular, Fisher et al. (2003) estimate that the reduction in "unwarranted" variability of care can lead to a substantial decrease in the volume of services delivered, and an up to 30% reduction in Medicare expenses, without affecting the quality of or access to care.

Of course, these "price" and "volume" metrics are connected, and with the adoption of the ACA a series of approaches designed to affect both have been gaining popularity. Payers (Medicare as well as private insurers) are experimenting with a number of approaches focused on incentivizing

both patients and providers to reduce their use of unwarranted services. On the patient side, the proliferation of high-deductible insurance plans has increased the patient role in making informed care-related choices. On the provider side, the new approaches include encouraging the formation of accountable care organizations (ACOs) to ensure the coordination and continuity of care (PPACA 2010, Berwick 2011) and the introduction of "bundled" payments associated with "episodes of care" (BPCI 2016).

One of the key features of the rapidly changing incentive landscape that affects both patients and providers is the growing number of performance-based incentive programs that condition payments to providers on the quality of care delivered (for a detailed review of performance-based incentive programs in OECD countries see Cashin et al. 2014). Currently, one of the largest performance-based programs in the US is the hospital value-based purchasing (VBP) program authorized in the ACA and run by the Centers for Medicare and Medicaid Services (CMS) since 2012 (VBP 2016). Under the VBP program, the CMS withholds a percentage (currently 1.75%) of a "base" diagnostic-related group (DRG) payment under the Medicare program and redistributes the withheld funds among approximately $3,500$ participating hospitals based on their total performance scores (TPS). TPS are calculated using hospital performance in four categories: clinical process of care, clinical outcomes, patient experience, and efficiency measures. Within each category, the TPS for a particular hospital is calculated based on its performance relative to the median and the 95th percentile of performance levels across all hospitals, inducing hospitals to compete on service quality. At the same time, to further raise competitive pressure on hospitals, an increasing amount of performance and cost data is being collected and made available to the public through the Hospital Compare program to help patients make informed choices when selecting a care provider (Hospital Compare 2016).

The notion of "service quality" is central to any healthcare initiative that aims to align payment for services with desired performance targets. Healthcare service quality has multiple dimensions and, in the hospital context, includes both clinical outcomes and clinical process measures. For example, the clinical outcomes used by the Medicare VBP and Hospital Compare programs include 30-day mortality rates from acute myocardial infarction, heart failure, and pneumonia as well as complication and infection rates. On the process side, the quality metrics include adherence to evidence-based medical procedures (such as the prophylactic administration of antibiotics to surgical patients and their discontinuance within 24 hours of surgery or the continuation of pre-admission beta-blocker therapy for surgical patients during the perioperative period) as well as metrics characterizing timely access to and delivery of care (such as the initiation of fibrinolytic therapy within 30 minutes of hospital arrival for a patient with acute myocardial infarction or the average time patients spend in the emergency department before being seen by a healthcare

professional). Timely access to care, in particular, plays an important role in shaping care outcomes in both inpatient and outpatient settings. In the US, the adoption of the ACA has helped to substantially lower the barrier for patients to access care "in principle," shifting the emphasis to patients' accessing care "on time." Following a substantial increase in the number of patients who are eligible for care, the "timeliness" component of healthcare delivery is likely to become one of the front-and-center issues in the US, as it has been for some time in a number of OECD countries. Data on patient waiting times in the US are not yet collected in a comprehensive fashion, but the limited available evidence (Thomson et al. 2013) indicates that patients in the US may have to wait longer for a primary care or specialist appointment than, for example, those in the UK. As a recent report by Merritt Hawkins indicates, appointment waiting times for both primary and specialist care in 15 major metropolitan markets in the US can be substantial and show significant regional variation: for example, the average wait to see a physician (for five physician specialties) is around 18 days, ranging from around 10 days in Dallas to more than 45 days in Boston (Merritt Hawkins 2014).

An important example of performance incentives based on patient waiting time targets is provided by the UK's public National Heath Service (NHS). The NHS is a monopsonist buyer of healthcare services that keeps track of not only how soon the necessary care is delivered after a patient arrives at a hospital in an emergency or for a scheduled elective procedure but also how many weeks or months a patient must wait before receiving elective care. In 2004, the NHS introduced a series of waiting time targets that included an 18-week target for the maximum waiting time from referral by a primary care physician to hospital treatment and a four-hour target for a patient arriving in an emergency department to be treated and either admitted or discharged. These targets have become one of the key indicators of hospital performance reported to the public (Lewis and Appleby 2006). Based on these targets, a set of performance incentives was introduced in the service purchasing contracts signed between hospitals and the NHS. For example, the 2008–2009 NHS Standard Contract allowed the NHS to withhold 0.5% of revenue paid to a hospital for every percent of patients that had to wait more than 18 weeks to receive elective care up to a maximum reduction of 5% (Department of Health 2008). Propper et al. (2008b) and Propper et al. (2010) provide empirical evidence that the introduction of waiting time targets has led to a significant fall in patient waiting times without impacting other aspects of patient care. Along with the introduction of performance-based incentives, the NHS has also injected an element of competition between hospitals by ensuring that every patient requiring treatment has the choice of several hospitals from which to receive care (NHS Choice 2016).

In the rapidly changing landscape of the US healthcare system, the large payers (such as Medicare) apply new performance-based incentive schemes to providers that are facing increasing competition for patients. A growing number of urgent care centers as well as pharmacies compete with

both primary care providers and emergency rooms (Lee et al. 2013). Road signs advertise low waiting times at hospital emergency rooms (O'Reilly 2010), and TV channels are filled with direct-to-consumer hospital advertising (Schenker et al. 2014). This new competitive dynamic is largely based on the quality of care provided rather than the cost to a patient. This is not surprising since the pricing of healthcare services in the US remains complex and is often not transparent, with the cost of care often not known to a patient until some time after the care has been delivered (HFMA 2014). As a reaction to the increase in competitive pressure, care providers are undergoing consolidation at record rates, which, in turn, prompts the consolidation of private insurance companies (Vaida and Weiss 2015). Despite this provider consolidation, however, competition for patients is likely to remain a significant factor that, along with performance-based incentives, influences the process of care delivery and health outcomes.

The simultaneous presence of competition and performance-based incentives in healthcare markets motivated us to focus on a set of questions about the nature of interaction between these two factors. More specifically, in this paper we analyze how a payer for healthcare services, such as a government agency (Medicare in the US or the NHS in the UK), should design performance-based contracts for providers (hospitals) that use service quality to compete for patients. We look at the setting where a payer uses patient access to care as a measure of service quality and aims to achieve a certain level of service quality as measured by the expected time patients must wait to receive care. In order to achieve this level of access, the payer imposes a compensation contract on all care providers that ties hospital compensation to the level of access a hospital delivers to its patients. Following the service operations management literature (e.g. Allon and Federgruen 2007, Allon and Federgruen 2008), we model the patient service dynamics at each hospital as that of an $M/M/1$ queue. This assumption is often used in the literature due to the analytical tractability it generates. While being decidedly simplistic, this assumption, in our opinion, still adequately captures the uncertain nature of demand for hospital services and the uncertain load such demand places on hospital resources in a qualitative manner. Similar to Allon and Federgruen (2008), we assume that the patient demand rate for a particular hospital is a function of its own service level as well as the service levels of its competitors. On the side of service capacity, we assume that each hospital faces an increasing convex cost of providing service capacity and determines its capacity level to maximize its expected profit. The payer influences hospitals' capacity decisions through a performance-based payment contract that is designed to minimize the payer's cost of achieving a certain service-level target.

We begin our analysis by looking at the hospital Nash equilibrium capacity decisions under a fixed performance-based incentive scheme represented by a payment function monotone in hospital service levels. In particular, we identify a set of regularity and sufficient conditions for the existence

of a unique set of Nash equilibrium hospital service levels. In our model, we assume that hospital demand rates are submodular functions of service levels. This assumption, combined with the convexity of hospital capacity costs, allows for hospital profit functions that are neither supermodular nor submodular in service levels. This feature of the service-level game that we study complicates the equilibrium analyses compared to the supermodular games considered in Allon and Federgruen (2007) and Allon and Federgruen (2008). Nevertheless, we identify the sufficient conditions for performance incentives to improve service levels as well as the conditions, such as the use of overly aggressive service-level targets, under which service levels deteriorate upon the introduction of performance incentives. Similarly, we provide an analytical description of settings where provider competition is beneficial for patient service as well as where competition may reduce patient access to care.

On the payer's side, we provide an analysis of the optimal contracting problem for the special case of a duopoly where the cost of capacity for each competing hospital is quadratic in hospital service rate and the performance-based hospital compensation functions and demand rates are linear in the hospital service level.

To the best of our knowledge, our paper is the first to study the performance-based contracting problem in a healthcare market where there is competition on service levels. The performance indicators, such as waiting time targets, used in performance-based programs in healthcare settings have not yet been the focus of otherwise extensive theoretical economics and regulation literature on incentive design (see, e.g. Chalkley and Malcomson 1998, Laffont and Tirole 1993, De Fraja 2000). The operations management literature, on the other hand, contains a number of papers that focus on performance-based incentives in services in general (Akan et al. 2011) as well as in call centers (Ren and Zhou 2008, Hasija et al. 2008) and in healthcare settings (So and Tang 2000, Fuloria and Zenios 2001, Jiang et al. 2012, Lee and Zenios 2012, Andritsos and Aflaki 2015). These papers, with the exception of Andritsos and Aflaki (2015), analyze settings with a monopolistic service provider. In Andritsos and Aflaki (2015), a comparison between the monopolistic and two duopolistic settings reveals that the introduction of competition can hamper a hospital's ability to achieve economies of scale and can also increase waiting times. Our paper focuses on a more complex setting where competition between providers occurs in the presence of performance-based incentives imposed by a payer that is focused on achieving a certain service-level target.

In the health economics literature, a significant number of studies focus on the effects of competition on the quality of care delivered (see Gaynor and Town 2012 for a detailed review). On the empirical side, the evidence on the impact of competition on the quality of care is nuanced. Cooper et al. (2011), Bloom et al. (2011), and Gaynor et al. (2013) provide evidence that the introduction of non-price competition in the NHS in 2000s led to improved patient choices and increased quality

of care, as measured by outcomes such as survival rates and patient waiting times. Propper et al. (2004) and Propper et al. (2008a), on the other hand, show that the price-based competition in the NHS internal market in 1990s reduced the quality of care (i.e. increased hospital mortality rates) as well as increased patient waiting times. On the theoretical side, Brekke et al. (2008) developed a model that describes competition between hospitals in the presence of "high-benefit" patients, who choose between hospitals based on waiting times and travel costs, and "low-benefit" patients, who go to the nearest hospital. It has been shown, in particular, that inter-hospital competition may lead to longer waits if the "high-benefit" patient segment is sufficiently large. Note that Brekke et al. (2008) do not model uncertainty in care demand or supply or the impact of performance-based incentives.

In the operations management domain, there is substantial literature on price- and service-level competition (see, e.g. Cachon and Harker 2002, Bernstein and Federgruen 2004, Allon and Federgruen 2007, Allon and Federgruen 2008, Allon and Federgruen 2009). Our hospital service modeling follows Allon and Federgruen (2007) in that we use $M/M/1$ queueing dynamics to describe hospital service dynamics and define the reciprocals of expected patient waiting time as measures of hospital service levels. Our analysis, however, has two distinguishing features that reflect the unique reality of healthcare markets. First, we focus on non-price competition and include performance-based incentives as an important factor governing the service-level equilibrium. Second, our study of the impact of competition includes the analysis of settings with varying degrees of competitive pressure, i.e. it presents a comparison of settings with different numbers of competitors.

The rest of the paper is organized as follows. In the next section, we introduce our model and a number of assumptions related to patient demand and provider cost structure. In Section 3, we analyze the hospital service level-selection problem in both monopolistic and competitive settings and investigate the effect of competition and performance-based incentives on hospital service levels. In Section 4, we look at the Nash equilibrium service levels and optimal performance contract parameters in a special duopoly setting. Finally, we discuss our results in Section 5.

2. The Model

We consider a healthcare system consisting of a single payer (for example, a government agency) and n competing medical facilities (hospitals), indexed by $i = 1, ..., n$. In such a system, the payer acts as a Stackelberg leader by using a performance-based contract to induce hospital investment in patient service capacity in order to improve patient access to care.

Actual patient service dynamics in a hospital are complex, and in our analysis we focus on a simplistic model frequently used in the literature (see, for example, Allon and Federgruen 2008

and references therein) that treats the patient care dynamic in each hospital as that in an $M/M/1$ queue. We assume that patients form a single group with homogeneous treatment needs and service times. The level of simplification in our model reflects a "macro" view of patient service that ignores a number of operational details, but this also allows us to gain insight into the effect of performance-based incentives on the time patients have to wait before receiving care. We denote the expected patient waiting time at hospital i by w_i, $i=1,...,n$ and, following Allon and Federgruen (2008), use

$$\theta_i = \frac{1}{w_i}, i=1,...,n \tag{1}$$

as a set of "service levels" that drive patient demand for hospital services. We note that the waiting times we model are designed to reflect the appointment delay patients experience before receiving care, i.e. the delay between the time a patient joins a queue (schedules her appointment) and the time her service is completed.

We assume that the service-level value for hospital i, θ_i is non-negative and limited from above by the upper bound $\bar{\theta}_i$, representing the highest service level the hospital can provide. We let $\lambda_i^{(n)}(\boldsymbol{\theta})$ denote the daily patient demand for hospital i in a setting with n hospitals on the market delivering service levels $\boldsymbol{\theta}=(\theta_1,...,\theta_n)$, $\boldsymbol{\theta}\in\boldsymbol{\Theta}$, where $\boldsymbol{\Theta}=[0,\overline{\theta}_1]\times...\times[0,\overline{\theta}_n]$. In what follows we also use the notation $\boldsymbol{\theta}_{-i}$ to designate the $(n-1)$-dimensional vector obtained from $\boldsymbol{\theta}$ by dropping θ_i and $(\boldsymbol{\theta}_{-i},\theta_i)$ as an alternative designation for $\boldsymbol{\theta}$. Similar to $\boldsymbol{\Theta}$, we use $\boldsymbol{\Theta}_{-i}$ to denote $[0,\overline{\theta}_1]\times...\times[0,\overline{\theta}_{i-1}]\times[0,\overline{\theta}_{i+1}]\times...\times[0,\overline{\theta}_n]$. Finally, $\lambda_i^{(1)}(\theta_i)$ designates the demand rate for a monopolistic hospital i.

We make the following assumptions about the general shape of the demand function for each hospital $i=1,...,n$ on $\boldsymbol{\Theta}$.

ASSUMPTION 1. *$\lambda_i^{(n)}(\boldsymbol{\theta})$ is nonnegative and $\lambda_i^{(n)}(\boldsymbol{\theta})=0$ if and only if $\theta_i=0$.*

ASSUMPTION 2. *$\lambda_i^{(n)}(\boldsymbol{\theta})$ is strictly increasing in θ_i and is strictly decreasing in θ_j for $\theta_i>0$:*

$$\frac{\partial\lambda_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_i} > 0, \tag{2}$$

$$\frac{\partial\lambda_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_j} < 0, \theta_i>0, j\neq i, j=1,...,n. \tag{3}$$

ASSUMPTION 3. *$\lambda_i^{(n)}(\boldsymbol{\theta})$ is twice continuously differentiable, strictly concave in θ_i, and submodular in θ_i and θ_j:*

$$\frac{\partial^2\lambda_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_i^2} < 0, \tag{4}$$

$$\frac{\partial^2\lambda_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j} < 0, j\neq i, j=1,...,n. \tag{5}$$

Assumption 4.

$$\frac{\partial \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i} > \sum_{j \neq i} \left| \frac{\partial \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_j} \right|. \tag{6}$$

Assumption 5.

$$\lambda_i^{(n-1)}\left(\boldsymbol{\theta}_{-j}\right) = \lambda_i^{(n)}\left(\boldsymbol{\theta}_{-j}, 0\right), j \neq i, j = 1, ..., n. \tag{7}$$

We consider Assumptions 1–3 to be intuitive. In particular, Assumption 1 indicates that patients do not patronize a hospital that does not provide a service. The second-order properties of the demand functions expressed by Assumption 3 state that a hospital's ability to attract new patients by improving service it provides declines with its own service level as well as the service levels of its competitors.

The term "diagonal dominance" is often associated with Assumption 4. This Assumption indicates that the impact of an increase in a hospital's service level on its demand outweighs the combined effect of a similar increase in the service levels of its competitors. Such an assumption is commonly used in the game-theoretical literature (see, e.g. Kolstad and Mathiesen 1987, Vives 2001, Bernstein and Federgruen 2004, Allon and Federgruen 2007, Allon and Federgruen 2008, Allon and Federgruen 2009). Similarly to these papers, we require Assumption 4 to prove the uniqueness of the Nash equilibrium arising in a market of competing hospitals.

Finally, Assumption 5 is necessary in order to study the effects of a hospital's exit from the market on the remaining competitors. In particular, this assumption states that upon a hospital's market exit, the demand functions for the remaining competitors can be evaluated by setting the service level of the exiting hospital to 0.

For notational convenience, we define the following non-negative parameters for all $i, j \neq i$:

$$\gamma_{ii} = \frac{\partial \lambda_i^{(n)}\left(\bar{\theta}_1, ..., \bar{\theta}_n\right)}{\partial \theta_i}, \tag{8}$$

$$\bar{\gamma}_{ii} = \frac{\partial \lambda_i^{(n)}\left(0, ..., 0\right)}{\partial \theta_i}, \tag{9}$$

$$\gamma_{ij} = -\min_{\boldsymbol{\theta}_{-i} \in \boldsymbol{\Theta}_{-i}} \frac{\partial \lambda_i^{(n)}\left(\boldsymbol{\theta}_{-i}, \bar{\theta}_i\right)}{\partial \theta_j}, \tag{10}$$

$$\bar{\gamma}_{ij} = -\max_{\boldsymbol{\theta}_{-i} \in \boldsymbol{\Theta}_{-i}} \frac{\partial \lambda_i^{(n)}\left(\boldsymbol{\theta}_{-i}, 0\right)}{\partial \theta_j}, \tag{11}$$

$$\delta_{ii} = -\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \frac{\partial^2 \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i^2}, \tag{12}$$

$$\delta_{ij} = -\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \frac{\partial^2 \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}, \tag{13}$$

$$\bar{\delta}_{ij} = -\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \frac{\partial^2 \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}. \tag{14}$$

Here, γ_{ii} and $\bar{\gamma}_{ii}$ reflect the minimum and maximum "degrees," respectively, of the monotonicity of the arrival rate for hospital i, λ_i, with respect to its own service level θ_i, while γ_{ij} and $\bar{\gamma}_{ij}$ reflect the minimum and maximum "degrees," respectively, of the monotonicity of the arrival rate for hospital i, λ_i, with respect to the service level of its j-th competitor. Furthermore, δ_{ii} indicates the minimum "degree" of concavity of λ_i with respect to θ_i, and δ_{ij} and $\bar{\delta}_{ij}$ indicate the minimum and maximum "degrees," respectively, of submodularity of λ_i with respect to (θ_i, θ_j), $j \neq i$.

In terms of supplying service capacity, we assume that each hospital i can set its daily patient service rate, μ_i, at a cost $C_i(\mu_i)$. Note that the stability of the resulting queue requires that $\mu_i > \lambda_i^{(n)}(\boldsymbol{\theta})$. We make an assumption that the hospital capacity cost functions are increasing and convex (Allon and Federgruen 2009, Brekke et al. 2008):

ASSUMPTION 6. *$C_i(\mu_i)$ is twice continuously differentiable and*

$$\frac{\partial C_i}{\partial \mu_i} > 0, \frac{\partial^2 C_i}{\partial \mu_i^2} \geq 0, i = 1, ..., n. \tag{15}$$

Note that under $M/M/1$ patient service dynamics, the expected patient waiting/sojourn time is given by $w_i = \frac{1}{\mu_i - \lambda_i^{(n)}(\boldsymbol{\theta})}$, so that $\mu_i = \lambda_i^{(n)}(\boldsymbol{\theta}) + \theta_i$. Thus, the service rates for all hospitals are determined by the set of their service levels θ_i. Similar to the patient demand functions, we introduce the following positive constants to characterize the hospitals' cost functions:

$$c_i = C_i'(0), \tag{16}$$

$$\bar{c}_i = C_i'\left(\lambda_i^{(1)}\left(\bar{\theta}_i\right) + \bar{\theta}_i\right), \tag{17}$$

$$d_i = \min_{0 \leq \mu_i \leq \lambda_i^{(1)}\left(\bar{\theta}_i\right) + \bar{\theta}_i} C_i''(\mu_i). \tag{18}$$

In (16) and (17), c_i and $\bar{c}_i$ are the minimum and maximum "degrees" of the monotonicity of the cost function with respect to the service rate, respectively, and d_i in (18) is the minimum "degree" of convexity of the cost function with respect to the service rate.

The performance-based compensation contract focuses on incentivizing hospitals to deliver a certain level of patient service as measured by the expected time patients have to wait before receiving care. Such incentives can be implemented via "sliding-scale" compensation $R_i(\theta_i, T, \boldsymbol{\xi})$ that a hospital i operating at the service level θ_i receives for serving a patient, with $T \geq T_l > 0$ describing the expected waiting time target with a positive lower bound and $\boldsymbol{\xi}$ describing the set of other performance parameters imposed on all competing hospitals. Since $R_i(\theta_i, T, \boldsymbol{\xi})$ must be a non-decreasing, non-negative function of θ_i on $[0, \bar{\theta}_i]$ for any choice of T and $\boldsymbol{\xi}$, we use

$$\boldsymbol{\Xi}(T) = (\boldsymbol{\xi} | R_i(0, T, \boldsymbol{\xi}) \geq 0) \tag{19}$$

to denote the set of feasible values of $\boldsymbol{\xi}$ for given $T \geq T_l$.

One example of such a compensation scheme is the "standard contract" used by the NHS in the UK (Department of Health 2008). Under this contract, a penalty is imposed on a hospital when the delay for patients to access hospital services exceeds a threshold of $T = 18$ weeks. The overall penalty is calculated as a product of the penalty rate $\eta = 0.5$, the per-patient "base" revenue, and the number of patients who had to wait longer than 18 weeks to be treated. Under $M/M/1$ patient service dynamics, the expected patient waiting time is given by $w_i = \frac{1}{\mu_i - \lambda_i^{(n)}(\boldsymbol{\theta})}$ and the fraction of patients served at hospital i whose wait exceeds the target $T > 0$ is given by $e^{-\frac{T}{w_i}} = e^{-\theta_i T}$. Thus, on a per-patient basis, if r_i is the "base" revenue a hospital receives, the "sliding-scale" compensation function under such a contract can be expressed as $R(\theta_i, T, \eta) = r_i\left(1 - \eta e^{-\theta_i T}\right)$. In this example, $\boldsymbol{\xi} = (\eta)$, $\boldsymbol{\Xi}(T) = [0, 1]$ and R_i is a monotone increasing, concave function of θ_i and a monotone increasing function of T.

In our analysis, we consider the compensation functions $R_i(\theta_i, T, \boldsymbol{\xi})$ that have the same properties as the functions in the NHS example:

ASSUMPTION 7. *$R_i(\theta_i, T, \boldsymbol{\xi})$ is twice continuously differentiable in θ_i on $[0, \bar{\theta}_i]$ and is continuously differentiable in T on $[T_l, +\infty)$ and*

$$\frac{\partial R_i}{\partial \theta_i} \geq 0, \frac{\partial^2 R_i}{\partial \theta_i^2} \leq 0, \frac{\partial R_i}{\partial T} \geq 0, i = 1, ..., n, \forall \boldsymbol{\xi} \in \boldsymbol{\Xi}(T). \tag{20}$$

Based on Assumption 7, we use the following values to describe the fee function for any $T \in [T_l, +\infty)$ and $\boldsymbol{\xi} \in \boldsymbol{\Xi}(T)$:

$$\underline{R}_i(T, \boldsymbol{\xi}) = R_i(0, T, \boldsymbol{\xi}), \tag{21}$$

$$\bar{R}_i(T, \boldsymbol{\xi}) = R_i\left(\bar{\theta}_i, T, \boldsymbol{\xi}\right), \tag{22}$$

$$\underline{R}_i'(T, \boldsymbol{\xi}) = \frac{\partial R_i}{\partial \theta_i}\left(\bar{\theta}_i, T, \boldsymbol{\xi}\right), \tag{23}$$

$$\bar{R}_i'(T, \boldsymbol{\xi}) = \frac{\partial R_i}{\partial \theta_i}(0, T, \boldsymbol{\xi}). \tag{24}$$

In the following section, we use parameters (8)–(14), (16)–(18), and (21)–(24) to characterize the sufficient conditions for the existence and uniqueness of Nash equilibrium service levels for competing hospitals.

Given the structure of the performance-based contract, demand model, and service capacity cost function, we treat each hospital as a risk-neutral profit-maximizing entity. The choice of profit as a hospital objective is justified in many practical settings. For example, in the UK, NHS hospitals are funded by the public; however, they have substantial managerial and financial flexibility. In particular, hospitals can retain surplus cash and sell property and retain cash from any sale. As a result, hospitals are often described as profit maximizers (De Fraja 2000, Miraldo et al. 2011).

Note that the general form of the fee function R_i also allows us to model the behavior of hospitals that, in addition to profit, assign some altruistic, non-monetary value to treating patients (Brekke et al. 2008).

In our model, for any combination of performance-based parameters $T \in [T_l, +\infty)$ and $\boldsymbol{\xi} \in \boldsymbol{\Xi}(T)$ and given set of competitors' service levels $\boldsymbol{\theta}_{-i}$, the optimal response problem for hospital i is expressed as

$$\max_{\theta_i} \ \pi_i^{(n)}(\boldsymbol{\theta}, T, \boldsymbol{\xi}) \triangleq \lambda_i^{(n)}(\boldsymbol{\theta}) R_i(\theta_i, T, \boldsymbol{\xi}) - C_i\left(\lambda_i^{(n)}(\boldsymbol{\theta}) + \theta_i\right) \tag{25}$$

$$\text{s.t.} \ \ 0 \le \theta_i \le \bar{\theta}_i. \tag{26}$$

In Section 3 we describe the set of sufficient conditions that ensures the existence and uniqueness of the Nash equilibrium service levels $\boldsymbol{\theta}^{\mathrm{NE}}(T, \boldsymbol{\xi}) = (\theta_1^{\mathrm{NE}}(T, \boldsymbol{\xi}), ..., \theta_n^{\mathrm{NE}}(T, \boldsymbol{\xi}))$ for the problems defined by (25) and (26) for each $i = 1, ..., n$.

We assume that the risk-neutral payer selects the performance-based contract parameters T and $\boldsymbol{\xi}$ to minimize the overall expected payment required to achieve a certain level of patient access to care. Specifically, the payer's optimization problem can be expressed as

$$\min_{T, \boldsymbol{\xi}} \ \sum_{i=1}^{n} \lambda_i^{(n)}\left(\boldsymbol{\theta}^{\mathrm{NE}}(T, \boldsymbol{\xi})\right) R_i\left(\boldsymbol{\theta}^{\mathrm{NE}}(T, \boldsymbol{\xi}), T, \boldsymbol{\xi}\right) \tag{27}$$

$$\text{s.t.} \ \ \theta_i^{\mathrm{NE}}(T, \boldsymbol{\xi}) \ge \frac{1}{T}, i = 1, ..., n, \tag{28}$$

$$\pi_i^{(n)}\left(\boldsymbol{\theta}^{\mathrm{NE}}(T, \boldsymbol{\xi})\right) \ge 0, i = 1, ..., n, \tag{29}$$

$$T_l \le T \le T_h, \boldsymbol{\xi} \in \boldsymbol{\Xi}(T). \tag{30}$$

The objective function (27) represents the expected daily compensation that the payer distributes to all hospitals, while the constraint (28) ensures that patients do not have to wait more than T on average before receiving care at any hospital. The "participation" constraint (29) ensures that hospitals cannot lose money under the imposed performance-based contract. The upper bound T_h on the value of the waiting time target T reflects the maximum clinically acceptable expected duration a patient must wait before receiving care, while the lower bound $T_l > 0$ reflects the highest level of patient service that a payer can realistically request from a hospital.

3. Nash Equilibrium Analysis: Incentives and Competition

In this section we formulate sufficient conditions that ensure the existence and uniqueness of the service-level Nash equilibrium for fixed values of performance-based parameters $T \in [T_l, T_h)$ and $\boldsymbol{\xi} \in \boldsymbol{\Xi}(T)$. We also analyze how the equilibrium service levels that emerge in the presence of performance-based incentives are affected by the competitive environment and by the overall strength of the pressure of incentives on competitors' revenues.

3.1. Nash Equilibrium Service Levels: Existence and Uniqueness

Consider a setting described by (25)–(26) where a performance-based contract described by the fee function $R_i(\theta_i, T, \boldsymbol{\xi})$ is applied to each of n competing hospitals. To begin with, we are interested in identifying conditions that guarantee the existence of a unique set of Nash equilibrium service levels under this performance-based contract.

PROPOSITION 1. *Suppose that*

$$\underline{R}_i(T, \boldsymbol{\xi}) > \bar{c}_i, i = 1, ..., n \tag{31}$$

and

$$\bar{R}_i'(T, \boldsymbol{\xi}) < \frac{d_i}{3} - \frac{\left(-\delta_{ii} + \sum_{j \neq i} \delta_{ij}\right)^+ \left(\bar{R}_i(T, \boldsymbol{\xi}) - c_i\right)}{3\gamma_{ii}}, i = 1, ..., n, \tag{32}$$

where $x^+ = \max(x, 0)$. *Then, there exists a unique set of Nash equilibrium service levels for competing hospitals.*

Proposition 1 assumes that the per-patient revenue that hospitals receive from the payer is higher than hospitals' marginal costs for any service level they choose to provide (this assumption is expressed by (31)) and specifies an upper bound on the "strength" of the performance-based incentives (as expressed by the value of the marginal revenue improvement rate) that ensures that the service-level competitive dynamics is "well-behaved," i.e. that service-level competition between hospitals results in a unique equilibrium. In the absence of performance-based incentives ($\bar{R}_i' = 0, i = 1, ..., n$) the sufficient condition (32) holds; specifically, if $\delta_{ii} > \sum_{j \neq i} \delta_{ij}$, $i = 1, ..., n$, a condition that implies the second-order "diagonal dominance" property of the arrival rates for all hospitals:

$$\left| \frac{\partial^2 \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i^2} \right| > \sum_{j \neq i} \left| \frac{\partial^2 \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right|, \boldsymbol{\theta} \in \boldsymbol{\Theta}, i = 1, ..., n. \tag{33}$$

Conditions similar to (33) are often used in oligopoly pricing literature to ensure the uniqueness of the Nash equilibrium (see, for example, Vives 2001). In this regard, the sufficient condition (32) is less restrictive than that implied by (33).

In the example of the performance-based contract used by the NHS in the UK, the hospital compensation function is $R_i(\theta_i, T, \eta) = r_i - \eta r_i e^{-\theta_i T}$. Under this contract, the sufficient conditions (31) and (32) can be expressed as

$$\eta < \eta_i^\star(T) = \min\left(1 - \frac{\bar{c}_i}{r_i}, \left(\frac{1}{3T}\left[\frac{d_i}{r_i} - \frac{\left[-\delta_{ii} + \sum_{j \neq i} \delta_{ij}\right]^+ \left[1 - \frac{c_i}{r_i}\right]}{\gamma_{ii}}\right]\right)\right), i = 1, ..., n. \tag{34}$$

Thus, in this case, for each value of the expected waiting time target T Proposition 1 imposes the maximum value of the penalty rate η that guarantees the existence and uniqueness of the Nash equilibrium.

3.2. Effect of Incentives on Nash Equilibrium Service Levels

The main rationale for a payer to use performance-based incentives is to improve the level of service patients receive. While it is reasonable to assume that the introduction of incentives that discourage excessive patient waiting times should increase the service levels provided by competing hospitals, the actual effect of performance-based compensation depends on the shape of the compensation function and its relation to the compensation level hospitals received prior to the introduction of incentives. Let $R_{i,\mathrm{b}}$ be the "base" compensation value provided by the payer to all competitors in the absence of performance-based incentives and $R_i(\theta_i, T, \boldsymbol{\xi})$ be the compensation function applied to any competitor i, $i = 1, ..., n$ under the performance-based contract defined by the expected waiting time target T and other contract parameters $\boldsymbol{\xi} \in \boldsymbol{\Xi}(T)$. We assume that the introduction of incentives cannot lead to compensation that is higher than or lower than $R_{i,\mathrm{b}}$ for all feasible values of service level for hospital $i = 1, ..., n$. In other words, we assume that $\underline{R_i}(T, \boldsymbol{\xi}) \leq R_{i,\mathrm{b}} \leq \bar{R}_i(T, \boldsymbol{\xi})$ for all $i = 1, .., n$.

Below, we look at the potential impact of performance-based incentives in a setting with identical competitors. For a competitive setting with n identical hospitals, define

$$\gamma = \gamma_{ii}, \bar{\gamma} = \bar{\gamma}_{ii}, \delta = \delta_{ii}, c = c_i, \bar{c} = \bar{c}_i, d = d_i, i = 1, ..., n, \tag{35}$$

$$\Gamma = \gamma_{ij}, \bar{\Gamma} = \bar{\gamma}_{ij}, \Delta = \delta_{ij}, \bar{\Delta} = \bar{\delta}_{ij}, i, j = 1, ..., n, j \neq i. \tag{36}$$

On the compensation side, we define "symmetric" versions of (21)–(24) as follows:

$$\bar{R}(T, \boldsymbol{\xi}) = \bar{R}_i(T, \boldsymbol{\xi}), \underline{R}(T, \boldsymbol{\xi}) = \underline{R}_i(T, \boldsymbol{\xi}), \bar{R}'(T, \boldsymbol{\xi}) = \bar{R}_i'(T, \boldsymbol{\xi}), \underline{R}'(T, \boldsymbol{\xi}) = \underline{R}_i'(T, \boldsymbol{\xi}), i = 1, ..., n. \tag{37}$$

The following proposition outlines the effect of introducing performance-based incentives into a symmetric competitive setting.

PROPOSITION 2. *Suppose that*

$$\underline{R}(T, \boldsymbol{\xi}) > \bar{c}, \tag{38}$$

$$\left(\bar{R}(T, \boldsymbol{\xi}) - c\right)\left(-\delta + (n-1)\bar{\Delta}\right)^+ < d\gamma, \tag{39}$$

$$\bar{R}'(T, \boldsymbol{\xi}) < \frac{d\gamma - (-\delta + (n-1)\Delta)^+ \left(\bar{R}(T, \boldsymbol{\xi}) - c\right)}{3\gamma}. \tag{40}$$

Then, there exist unique symmetric Nash equilibria $\boldsymbol{\theta}_{\mathrm{b}}^{(n)} = \left(\theta_{\mathrm{b}}^{(n)}, ..., \theta_{\mathrm{b}}^{(n)}\right)$ *in the absence of performance-based incentives and* $\boldsymbol{\theta}^{(n)} = \left(\theta^{(n)}, ..., \theta^{(n)}\right)$ *in the presence of performance-based incentives. In addition,* $\theta^{(n)} \geq \theta_{\mathrm{b}}^{(n)}$ *if* $R\left(\theta_{\mathrm{b}}^{(n)}, T, \boldsymbol{\xi}\right) \geq R_{\mathrm{b}}$*, or,*

$$\frac{\frac{\partial R\left(\theta_{\mathrm{b}}^{(n)}, T, \boldsymbol{\xi}\right)}{\partial \theta_i}}{R_{\mathrm{b}} - R\left(\theta_{\mathrm{b}}^{(n)}, T, \boldsymbol{\xi}\right)} \geq \frac{\frac{\partial \lambda^{(n)}\left(\theta_{\mathrm{b}}^{(n)}, ..., \theta_{\mathrm{b}}^{(n)}\right)}{\partial \theta_i}}{\lambda^{(n)}\left(\theta_{\mathrm{b}}^{(n)}, ..., \theta_{\mathrm{b}}^{(n)}\right)}, i = 1, ..., n, \tag{41}$$

and $\theta^{(n)} \leq \theta_{\mathrm{b}}^{(n)}$ *otherwise.*

As Proposition 2 indicates, the ability of the performance-based contract to induce improvements in the service levels provided by competing hospitals depends on the nature of the performance-based incentives. Consider a general, non-symmetric setting where the sufficient conditions for the existence and uniqueness of the Nash equilibrium hold in both the absence and presence of performance-based incentives. In order to rationalize the results of Proposition 2, consider the following definition:

DEFINITION 1. A contract $(T, \boldsymbol{\xi})$ is *penalty-based* (*bonus-based*) if, for all $i = 1, ..., n$, $R_i\left(\theta_{i,\mathrm{b}}^{(n)}, T, \boldsymbol{\xi}\right) < R_{\mathrm{i,b}}$ ($R_i\left(\theta_{i,\mathrm{b}}^{(n)}, T, \boldsymbol{\xi}\right) > R_{i,\mathrm{b}}$). A contract $(T, \boldsymbol{\xi})$ is *neutral* if, for all $i = 1, ..., n$, $R_i\left(\theta_{i,\mathrm{b}}^{(n)}, T, \boldsymbol{\xi}\right) = R_{i,\mathrm{b}}$.

This definition is rather intuitive: The introduction of a penalty-based (bonus-based) contract reduces (increases) the compensation for all competing hospitals if they keep their service levels unchanged. The neutral contract, on the other hand, leaves hospital compensation unchanged if hospitals maintain the Nash equilibrium service levels attained in the absence of incentives. Proposition 2 states that in a symmetric setting, neither bonus-based nor neutral contracts can result in a decrease in patient service levels, while penalty-based contracts may indeed reduce the resulting service levels.

As an example, consider the performance-based contract (T, η) used by the NHS. Under such a contract, $R_{i,\mathrm{b}} = r_i$, and $R_i(\theta_i, T, \eta) = R_{i,\mathrm{b}}(1 - \eta \exp(-\theta_i T)) < R_{i,\mathrm{b}}$. As such, it is a penalty-based contract. For such a contract, assuming that (34) holds, note that

$$\frac{\frac{\partial R_i\left(\theta_{i,\mathrm{b}}^{(n)}, T, \boldsymbol{\xi}\right)}{\partial \theta_i}}{R_{i,\mathrm{b}} - R_i\left(\theta_{i,\mathrm{b}}^{(n)}, T, \boldsymbol{\xi}\right)} = T \tag{42}$$

and that the service level in a symmetric setting can increase upon the introduction of the contract if T is sufficiently high, i.e. if

$$T \geq T^*, \tag{43}$$

where

$$T^* = \frac{\frac{\partial \lambda^{(n)}\left(\theta_{\mathrm{b}}^{(n)}, ..., \theta_{\mathrm{b}}^{(n)}\right)}{\partial \theta_i}}{\lambda^{(n)}\left(\theta_{\mathrm{b}}^{(n)}, ..., \theta_{\mathrm{b}}^{(n)}\right)}. \tag{44}$$

For this contract, the left-hand side of (41) is the reciprocal of the service-level target that such a contract aims to achieve. Thus, the payer must select moderate service-level targets for the incentive contract to result in an increase in patient service levels. Note that for a general "penalty-based" contract, the higher the difference in the numerator of the expression on the left-hand side of (41), the more aggressive the service-level target that such a contract aims to achieve.

The specific form of the compensation function used in this contract allows for a more general result on the monotonicity of the Nash equilibrium service levels with respect to the penalty rate

parameter η. Consider a setting where an NHS-type performance-based contract characterized by parameters η and T is applied to a setting with n identical hospitals characterized by (35)–(36), and let $\eta^\star(T) = \eta_i^\star(T)$, $i = 1, ..., n$ denote an upper bound from (34). Then, the monotonicity of the Nash equilibrium service levels can be characterized as follows.

PROPOSITION 3. *In a symmetric setting with n hospitals, let*

$$\bar{\eta}(T) = \frac{(\delta + (n-1)\bar{\Delta})\left(1 - \frac{\bar{c}}{r}\right) + \frac{d}{r}(\gamma + 1)}{(2\bar{\gamma} - (n-1)\bar{\Gamma})T + (\delta + (n-1)\bar{\Delta})}. \tag{45}$$

Suppose that

$$\eta < \min\left(\eta^\star(T), \bar{\eta}(T)\right). \tag{46}$$

Then, there exists a unique Nash equilibrium that is symmetric, $\boldsymbol{\theta}^{(n)}(T, \eta) = \left(\theta^{(n)}(T, \eta), ..., \theta^{(n)}(T, \eta)\right)$*, where* $\theta^{(n)}(T, \eta)$ *is a non-increasing (non-decreasing) function of* η *for* $T \leq T^\star$ *(*$T > T^\star$*), with* $T^\star$ *given by (44).*

As the statement for Proposition 3 indicates, the value of T may play a key role in shaping the effect of the NHS-type contract on competing hospitals. In particular, the low values of T may lead to counterproductive results in terms of the impact of the performance-based incentives: Stronger incentives can lead to lower patient service levels.

3.3. Effect of Competition on Nash Equilibrium Service Levels

The effect of competitive pressure on hospital service levels may vary depending on the interplay between the patient demand and capacity cost functions. In particular, as competitive pressure increases, hospitals may respond by improving patient service levels and reducing expected waiting times. However, it is also possible for the competition to result in increased waiting times.

The following proposition illustrates each of these outcomes for the case of two competing hospitals in the presence of a performance-based contract characterized by compensation functions $R_i(\theta_i, T, \boldsymbol{\xi})$, $i = 1, 2$. We assume that the compensation function satisfies (31) and (32) such that the unique Nash equilibrium exists. Specifically, let $\theta_1^{(1)}$ and $\theta_2^{(1)}$ be the optimal service levels of hospitals 1 and 2, respectively, in a monopolistic setting and let $\theta_1^{(2)}$ and $\theta_2^{(2)}$ be the Nash equilibrium service levels of hospitals 1 and 2, respectively, in a duopoly.

PROPOSITION 4. *(a) Suppose the cost functions* $C_i(\cdot)$ *are linear for both* $i = 1, 2$*. Then,*

$$\theta_i^{(2)} \leq \theta_i^{(1)}, i = 1, 2. \tag{47}$$

(b) Suppose that

$$d_i > \frac{\delta_{ij}}{\bar{\gamma}_{ij}(\gamma_{ii} + 1)}\left(\bar{R}_i(T, \boldsymbol{\xi}) - c_i\right), i, j = 1, 2, j \neq i \tag{48}$$

and

$$\bar{R}_i'(T,\boldsymbol{\xi}) < d_i(\gamma_{ii}+1) - \frac{\delta_{ij}}{\bar{\gamma}_{ij}}\left(\bar{R}_i(T,\boldsymbol{\xi}) - c_i\right), i,j=1,2, j\neq i. \tag{49}$$

Then,

$$\theta_i^{(2)} \geq \theta_i^{(1)}, i=1,2. \tag{50}$$

Proposition 4 indicates that the onset of competition may depress service levels in settings where the marginal costs of increasing provider capacity do not depend on the service level achieved. On the one hand, in a setting where the marginal costs of service improvement increase with the service rate ($d_i>0, i=1,2$), competitive pressure may lead to service level improvement provided that patients exhibit a substantial degree of sensitivity to the service levels they are subjected to, i.e. provided that $\gamma_{ii}, i=1,2$ are high enough.

Under the NHS-type performance-based contract, $R_i(\theta_i,T,\eta) = r_i - r_i\eta e^{-\theta_i T}$ and conditions (48) and (49) become $d_i \geq \frac{\delta_{ij}}{\bar{\gamma}_{ij}(\gamma_{ii}+1)}(r_i-c_i)$ and $\eta < \min_i\left(\frac{1}{r_i T}\left(d_i(\gamma_{ii}+1) - \frac{\delta_{ij}}{\bar{\gamma}_{ij}}(r_i-c_i)\right)\right)$, respectively. For this type of contract, the results of Proposition 4 can be extended, in part, beyond the duopoly setting. Consider an oligopoly setting with $n>2$ competing hospitals and let $\theta_i^{(n)}(\eta,T)$ be the Nash equilibrium service level of hospital i, $i=1,...,n$ in such oligopoly. By comparison, consider an oligopolistic setting with $n-1$ competing hospitals resulting from the n-hospital setting after hospital $j=1,...,n$ "exits" the market; let $\theta_i^{(n-1,j)}(\eta,T)$ be the Nash equilibrium service level of hospital $i\neq j$ in such an oligopoly with $n-1$ competing hospitals. The following proposition compares the Nash equilibrium service levels in these two settings.

PROPOSITION 5. *(a) Consider a symmetric oligopoly with $\gamma_{ii}=\gamma$, $\gamma_{ij}=\Gamma$, $\bar{\gamma}_{ii}=\bar{\gamma}$, $\bar{\gamma}_{ij}=\bar{\Gamma}$, $\delta_{ii}=\delta$, $\delta_{ij}=\Delta$, $\bar{\delta}_{ij}=\bar{\Delta}$, $c_i=c$, $\bar{c}_i=\bar{c}$, $d_i=d$, and $\eta^\star(T)=\eta_i^\star(T)$, $i,j=1,...,n$, $j\neq i$. Suppose that the hospitals' cost function $C(\cdot)$ is linear and $\eta<\eta^\star(T)$. Then,*

$$\theta_i^{(n)}(\eta,T) \leq \theta_i^{(n-1,j)}(\eta,T), i,j=1,...,n, j\neq i. \tag{51}$$

(b) Consider a general, non-symmetric oligopoly setting and let

$$D_i = \max_{j\neq i}\left(\frac{\delta_{ij}}{\bar{\gamma}_{ij}}\right), i=1,...,n, \tag{52}$$

$$\hat{\eta}_i(T) = \frac{1}{r_i T}\left(d_i(\gamma_{ii}+1) - D_i(r_i-c_i)\right), i,j=1,...,n, j\neq i. \tag{53}$$

Suppose that

$$d_i > \frac{D_i}{(\gamma_{ii}+1)}(r_i-c_i), i=1,...,n \tag{54}$$

and

$$\eta < \min_{i\in\{1,...,n\}}\left(\min\left(\eta_i^\star(T), \hat{\eta}_i(T)\right)\right). \tag{55}$$

Then,

$$\theta_i^{(n)}(\eta,T) \geq \theta_i^{(n-1,j)}(\eta,T), i,j=1,...,n, j\neq i. \tag{56}$$

Proposition 5 provides a more specific characterization of the effect of competition on service levels in the case of an NHS-type contract. In particular, it emphasizes the need for the penalty rate parameter η to stay low for the increase in the number of competitors in the presence of performance-based incentives to result in a decrease in expected patient waiting times.

In the following section we consider the example of a duopoly where the demands of competing hospitals are linear functions of their respective service levels and the costs are quadratic functions of their respective arguments. These specific forms of the demand and cost functions allow for sharper analytical characterizations of the Nash equilibrium service levels and optimal performance contract parameters.

4. Special Case: Duopoly with Linear Demand, Quadratic Cost, and Linear Compensation Functions

Consider a duopoly in which the demand functions for hospital $i=1,2$ are given by

$$\lambda_1^{(2)}(\theta_1,\theta_2) = \alpha_1\theta_1 - \rho_{12}\theta_1\theta_2, \tag{57}$$

$$\lambda_2^{(2)}(\theta_1,\theta_2) = \alpha_2\theta_2 - \rho_{21}\theta_1\theta_2, \tag{58}$$

the cost functions are

$$C_1(\theta_1,\theta_2) = b_1\left(\lambda_1(\theta_1,\theta_2)+\theta_1\right)^2, \tag{59}$$

$$C_2(\theta_1,\theta_2) = b_2\left(\lambda_2(\theta_1,\theta_2)+\theta_2\right)^2, \tag{60}$$

with $\alpha_i, b_i > 0$, $i=1,2$, and the "sliding-scale" compensation function is given by

$$R_i(\theta_i,T,\eta) = r_i\left(1+\eta\left(\theta_i-\frac{1}{T}\right)\right), \theta_i\in[0,\bar{\theta}_i], i=1,2, \tag{61}$$

where $T\geq\frac{1}{\bar{\theta}_i}, i=1,2$, $\boldsymbol{\xi}=\eta$, and $\boldsymbol{\Xi}(T)=[0,\min(1,T)]$. Note that (61) satisfies Assumption 7 and that Assumptions 1–6 are satisfied as long as

$$\alpha_i - \rho_{ij}\left(\bar{\theta}_i+\bar{\theta}_j\right) > 0, i,j=1,2, j\neq i. \tag{62}$$

Below, we conduct an analysis of the Nash equilibrium service levels in this setting.

4.1. Nash Equilibrium Service Levels: Analytical Characterization

In the presence of performance incentives, the service level with which hospital $i=1,2$ responds to the service level of its competitor $j=1,2, j\neq i$ is expressed by the following result.

LEMMA 1. *Assume that, in addition to (62),*

$$\eta < T. \tag{63}$$

Let θ_j, $j=1,2, j\neq i$ be the service level of the hospital competing with hospital $i=1,2$ and

$$\theta_i^{\mathrm{s}}(\theta_j) = \frac{(\alpha_i - \rho_{ij}\theta_j)(1-\eta/T)\left(\frac{r_i}{b_i}\right)}{2\left[\left(1+\alpha_i-\rho_{ij}\theta_j\right)^2 - (\alpha_i-\rho_{ij}\theta_j)\eta\left(\frac{r_i}{b_i}\right)\right]}. \tag{64}$$

Then, the optimal response for hospital i is to set its service level to

$$\theta_i^r = \min\left(\theta_i^{\mathrm{s}}, \bar{\theta}_i\right). \tag{65}$$

In Section 3 we showed that (31) and (32) ensure the existence and uniqueness of the Nash equilibrium service levels for general compensation, demand, and cost functions. Below, we provide sharper conditions for the case of a duopoly with compensation, demand, and cost functions given by (61), (57)–(58), and (59)–(60), respectively.

Proposition 6. *Suppose that, in addition to (62), the following conditions hold:*

$$\alpha_i > \left(\sqrt{\frac{\rho_{ij}r_i}{2b_i}} - 1\right)^+, i,j=1,2, j\neq i, \tag{66}$$

$$\eta < \eta^\star(T) = \min\left(T, (\alpha_i+1)\frac{b_i}{r_i}\right), i,j=1,2, j\neq i, \tag{67}$$

$$\bar{\theta}_i \leq \frac{1}{\rho_{ji}}\left(1+\alpha_j - \frac{\eta r_j}{2b_j} - \sqrt{\left(\frac{\eta r_j}{2b_j}\right)^2 + \left(\frac{\rho_{ji}r_j}{2b_j}\right)\left(1-\frac{\eta}{T}\right)}\right), i,j=1,2, j\neq i, \tag{68}$$

$$2\bar{\theta}_i + \bar{\theta}_j < \frac{1}{\rho_{ij}}\left(1+\alpha_i - \frac{\eta r_i}{b_i}\right), i,j=1,2, j\neq i. \tag{69}$$

Then, there exist unique Nash equilibrium service levels $\hat{\theta}_1^{\mathrm{NE}}(T,\eta)$ and $\hat{\theta}_2^{\mathrm{NE}}(T,\eta)$.

The results of Proposition 6 indicate that the competition between two hospitals produces a well-defined outcome in terms of service levels if several main conditions are satisfied. First, as (66) implies, patients' sensitivity to changes in their hospital's service level must be substantially stronger than their sensitivity to changes in a competitor's service level, beyond what is implied by (62). Second, as indicated by (67), the incentive fee parameter η is limited from above by a threshold $\eta^\star(T)$ that is increasing in the expected patient waiting time target T for small values of T and decreasing in T for large values of T. Finally, the service levels provided by competitors must be limited from above, which effectively limits the set of waiting times achievable by an incentive contract of the type described by (61).

In many settings, conditions (66)–(69) are less restrictive than the general conditions of Proposition 1. As an illustrative example, consider a symmetric setting with $r_1=r_2=r$, $b_1=b_2=b$, $\rho_{12}=\rho_{21}=\rho$, $\alpha_1=\alpha_2=\alpha$, and $\bar{\theta}_i=\bar{\theta}_j=\bar{\theta}$. Then, (66)–(69) become

$$\alpha > \left(\sqrt{\frac{\rho r}{2b}} - 1\right)^+, \tag{70}$$

$$\eta < \min\left(T, (1+\alpha)\frac{b}{r}\right), \tag{71}$$

$$\bar{\theta} \leq \min\left(\frac{1}{3\rho}\left(\alpha+1-\frac{\eta r}{b}\right), \frac{1+\alpha-\frac{\eta r}{2b}-\sqrt{\left(\frac{\eta r}{2b}\right)^2+\left(\frac{\rho r}{2b}\right)\left(1-\frac{\eta}{T}\right)}}{\rho}\right). \tag{72}$$

On the other hand, for the demand and cost functions (57)–(60) and fee function (61), the problem parameters in the conditions of Proposition 1 can be expressed as

$$\underline{R}_i(T,\boldsymbol{\xi}) = r\left(1-\frac{\eta}{T}\right), \tag{73}$$

$$\bar{c}_i = 2b(\alpha+1)^2\bar{\theta}, \tag{74}$$

$$\bar{R}'_i(T,\boldsymbol{\xi}) = r\eta, \tag{75}$$

$$d_i = 2b, \tag{76}$$

$$\delta_{ii} = 0, \tag{77}$$

$$\delta_{ij} = \rho, \tag{78}$$

$$\bar{R}_i(T,\boldsymbol{\xi}) = r\left(1+\eta\left(\bar{\theta}-\frac{1}{T}\right)\right), \tag{79}$$

$$c_i = 0, \tag{80}$$

$$\gamma_{ii} = \alpha-\rho\bar{\theta}, i,j=1,2, j\neq i. \tag{81}$$

Thus, (31) and (32) are equivalent to

$$r\left(1-\frac{\eta}{T}\right) > 2b(\alpha+1)^2\bar{\theta} \tag{82}$$

and

$$r\eta < \frac{2b}{3} - \frac{\rho}{3(\alpha-\rho\bar{\theta})} r\left(1+\eta\left(\bar{\theta}-\frac{1}{T}\right)\right). \tag{83}$$

Note that to both sets of conditions we must also add (62), or

$$\bar{\theta} < \frac{\alpha}{2\rho}. \tag{84}$$

In addition, the general conditions (82) and (83) must be augmented by $\eta < T$, which ensures that the compensation function is positive for any service-level value.

In the case of a monopoly ($\rho = 0$), the sufficient conditions of Proposition 6 reduce to

$$\eta < \min\left(T, (1+\alpha)\frac{b}{r}\right), \tag{85}$$

while the sufficient conditions of Proposition 1 become

$$\eta < \min\left(T, \frac{2b}{3r}\right), \tag{86}$$

$$\bar{\theta} \leq \frac{\frac{r}{2b}\left(1-\frac{\eta}{T}\right)}{(\alpha+1)^2}. \tag{87}$$

Note that the sufficient conditions (86)–(87) are "tighter" than (85); specifically, for each combination of parameters T, α, r, and b, the right-hand side of (85) is greater than or equal to the right-hand side of (86).

The optimal response expressions of Proposition 1 do not allow closed-form analytical characterization of the Nash equilibrium except in special cases. Below, we look at the Nash equilibrium service levels in the case of a symmetric duopoly with $r_i = r$, $b_i = b$, $\alpha_i = \alpha$, $\bar{\theta}_i = \bar{\theta}$, $\rho_{ij} = \rho$, $i, j = 1, 2$.

PROPOSITION 7. *Suppose that (62) and (66)–(68) hold and consider a duopoly setting with* $r_i = r$, $b_i = b$, $\alpha_i = \alpha$, $\bar{\theta}_i = \bar{\theta}$, $i = 1, 2$, *and* $\rho_{ij} = \rho$, $i, j = 1, 2$. *For* $b > 0$, *define*

$$\hat{u} = \frac{\rho r}{6b} - \left(\frac{\alpha+1}{3}\right)^2 + \left(\frac{\eta r}{3b}\right)\left(\frac{\alpha+1}{3} + 1 - \frac{\rho}{2T}\right) - \left(\frac{\eta r}{3b}\right)^2 \tag{88}$$

$$\begin{aligned}\hat{v} = {} & \frac{\rho r}{4b}\left(\frac{\alpha+1}{3} - 1\right) - \left(\frac{\alpha+1}{3}\right)^3, \\ & + \frac{3}{2}\left(\frac{\eta r}{3b}\right)\left(\frac{\rho r}{6b} + \frac{\rho}{2T} - \left(\frac{\alpha+1}{3}\right)\left(2 + \frac{\rho}{2T}\right) + \left(\frac{\alpha+1}{3}\right)^2\right) \\ & + \frac{3}{2}\left(\frac{\eta r}{3b}\right)^2\left(\left(1 - \frac{\rho}{2T}\right) + \left(\frac{\alpha+1}{3}\right)\right) - \left(\frac{\eta r}{3b}\right)^3,\end{aligned} \tag{89}$$

and

$$\hat{\theta}(\eta, T) = \frac{2}{\rho}\left(\frac{\alpha+1}{3}\right) - \frac{1}{3\rho}\left(\frac{\eta r}{b}\right) + \frac{1}{\rho}\left(\left(\sqrt{(\hat{u})^3 + (\hat{v})^2} + \hat{v}\right)^{\frac{1}{3}} - \left(\sqrt{(\hat{u})^3 + (\hat{v})^2} - \hat{v}\right)^{\frac{1}{3}}\right). \tag{90}$$

Then, the unique Nash equilibrium service level for the competing hospitals is given by

$$\theta^{\mathrm{NE}}(\eta, T) = \min\left(\hat{\theta}(\eta, T), \bar{\theta}\right). \tag{91}$$

As the results of Proposition 7 indicate, while the presence of competition and introduction of incentives can improve patient service levels, they may also have a detrimental effect.

In order to gain insight into the types of problem settings where competition and incentives combine to improve patient service as well as settings in which these factors are detrimental to patient service, we select a base-case parameter set that corresponds to the realistic values for patient service levels and demand rates in the monopolistic setting without incentives that we use as a reference. Note that, as implied by (64), a monopolistic hospital in the absence of incentives will set its service level at $\theta^{\mathrm{M}} = \frac{\alpha}{(1+\alpha)^2}\frac{r}{2b}$ when there is no upper bound on the service level values, with the resulting daily demand rate of $\alpha\theta^{\mathrm{M}} = \left(\frac{\alpha}{1+\alpha}\right)^2 \frac{r}{2b}$. For the base case, we select a setting with $\theta^{\mathrm{M}} = 0.01$ (corresponding to an expected patient waiting time for an appointment of 100 days) and $\alpha\theta^{\mathrm{M}} = 100$ patients per day and use the basic fee-for-service compensation rate of $r = 200$. These values result in $\alpha = 10000$ and $b = 0.9998$. Figure 1 shows how the service levels in a monopoly and

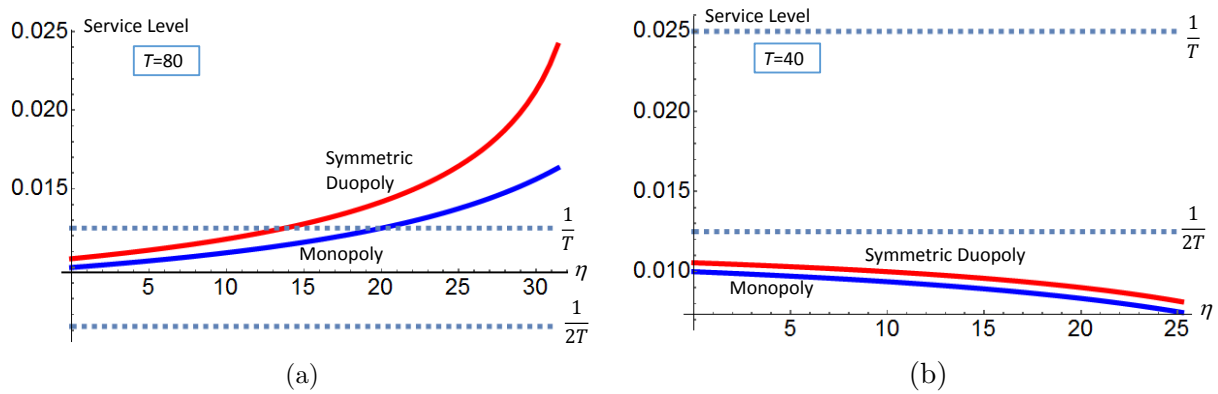

Figure 1 : The Nash equilibrium service levels in a monopoly and symmetric duopoly as functions of the incentive fee parameter η ($r=200$, $b=0.9998$, $\alpha=10000$, $\rho=50000$) for $T=80$ (a) and $T=40$ (b).

symmetric duopoly depend on the incentive fee parameter η for $T=80$ (Figure 1a) and $T=40$ (Figure 1b). In this figure, the values for η are limited to the range $[0,\eta^{\mathrm{cr}}(\alpha,\rho,r,b,T)]$, where

$$\eta^{\mathrm{cr}}(\alpha,\rho,r,b,T)=\min\left(\eta|\hat{\theta}(\alpha,\rho,r,b,\eta,T)=\bar{\theta}\right). \tag{92}$$

In addition, for each $\eta\in[0,\eta^{\mathrm{cr}}(\alpha,\rho,r,b,T)]$, $\bar{\theta}$ is set at

$$\bar{\theta}^{\max}(\alpha,\rho,r,b,T,\eta)=\min\left(\frac{1}{3\rho}\left(\alpha+1-\frac{\eta r}{b}\right),\frac{1+\alpha-\frac{\eta r}{2b}-\sqrt{\left(\frac{\eta r}{2b}\right)^2+\left(\frac{\rho r}{2b}\right)\left(1-\frac{\eta}{T}\right)}}{\rho}\right), \tag{93}$$

which is the right-hand side of (72).

Note that for $T=80$, both monopoly and duopoly service levels are increasing functions of η for $\eta<\eta^{\mathrm{cr}}(\alpha,\rho,r,b,T)$, in agreement with the results of Proposition 2. Indeed, for the problem setting we consider, (41) is equivalent to

$$\frac{1}{\frac{1}{T}-\theta_{\mathrm{b}}^{(n)}}\geq\frac{1}{\theta_{\mathrm{b}}^{(n)}}, \tag{94}$$

or

$$\theta_{\mathrm{b}}^{(n)}\geq\frac{1}{2T}, \tag{95}$$

where $\theta_{\mathrm{b}}^{(n)}=\hat{\theta}(\eta=0,T)$. (95) holds for $T=80$ for both $\rho=0$ (monopoly) and $\rho=50000$ (duopoly). As Figure 1a indicates, setting $\eta\approx13$ in a duopoly and $\eta\approx20$ in a monopolistic setting "lifts" the respective service levels to the target value $\frac{1}{T}$.

On the other hand, (95) is violated for $T=40$ for both $\rho=0$ and $\rho=50000$ and both monopoly and duopoly service levels are decreasing functions of η for $\eta<\eta^{\mathrm{cr}}(\alpha,\rho,r,b,T)$, as shown in Figure 1b. Thus, incentive contracts involving overly aggressive expected waiting time targets may actually result in a deterioration in patient waiting times.

As Figure 1 indicates, the base-case parameter set results in an increase in the service-level value when an identical competitor is added to a monopoly setting for all values of the incentive fee

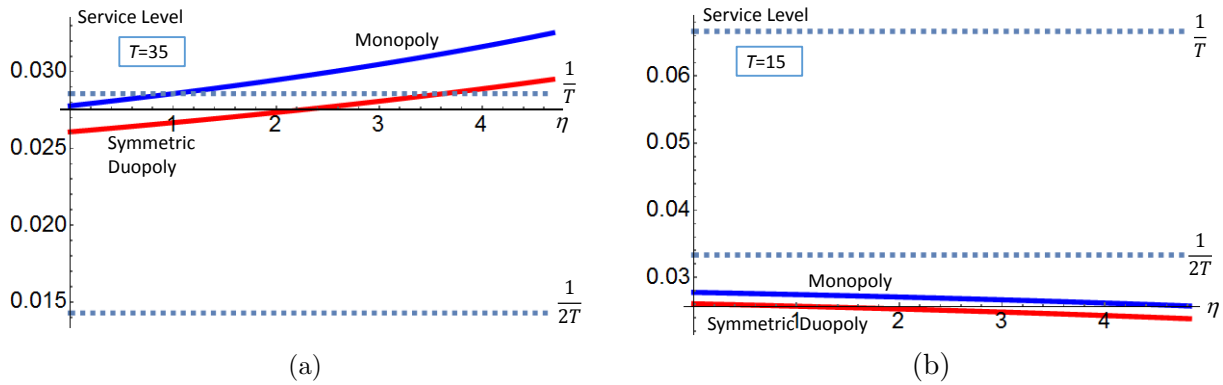

Figure 2 : The Nash equilibrium service levels in a monopoly and symmetric duopoly as functions of the incentive fee parameter η ($r=200$, $b=800$, $\alpha=0.5$, $\rho=3$) for $T=35$ (a) and $T=15$ (b).

parameter η below $\eta^{\mathrm{cr}}(\alpha,\rho,r,b,T)$. In order to understand why this is the case, note that, as follows from (64), in the absence of an upper bound on the service level values the monopoly service level is given by

$$\theta^{\mathrm{M}}=\frac{r}{2b}\left(1-\frac{\eta}{T}\right)\frac{\alpha}{(\alpha+1)^2-\left(\frac{\eta r}{b}\right)\alpha}, \tag{96}$$

while the corresponding Nash equilibrium symmetric duopoly service rate satisfies

$$\hat{\theta}=\frac{r}{2b}\left(1-\frac{\eta}{T}\right)\frac{\alpha-\rho\hat{\theta}}{\left(\alpha+1-\rho\hat{\theta}\right)^2-\left(\frac{\eta r}{b}\right)\left(\alpha-\rho\hat{\theta}\right)}. \tag{97}$$

The function

$$h(x)=\frac{x}{\left((x+1)^2-\left(\frac{\eta r}{b}\right)x\right)} \tag{98}$$

is monotone increasing in x for $0<x<1$ and monotone decreasing in x for $x>1$. Thus, for $\alpha-\rho\hat{\theta}>1$, $\hat{\theta}>\theta^{\mathrm{M}}$, as is the case in the examples in Figure 1. On the other hand, if $\rho\hat{\theta}<\alpha<1$, then the duopoly service level $\hat{\theta}$ is below the service level in monopoly θ^{M}. This last condition is achieved for any $\alpha<1$ and sufficiently low value of ρ. Figure 2 illustrates the case ($r=200$, $b=800$, $\alpha=0.5$, $\rho=3$) where the presence of competition leads to a decrease in the service rate. Note that, as in the base case, setting high performance targets in terms of expected waiting times may further decrease the service rate patients experience in a duopoly in the presence of performance-based incentives: While $T=35$ represents an achievable goal in both the monopoly and duopoly settings, $T=15$ is too aggressive to be achieved under the performance-based contract defined by (61).

In the examples shown in Figures 1 and 2, both the monopoly and duopoly service rates move in "unison" upon the introduction of performance-based incentives: They both either increase or decrease as functions of η. It is interesting to note that, as implied by Proposition 2, it is possible to have monopoly and duopoly service levels that "move" in the opposite directions under the

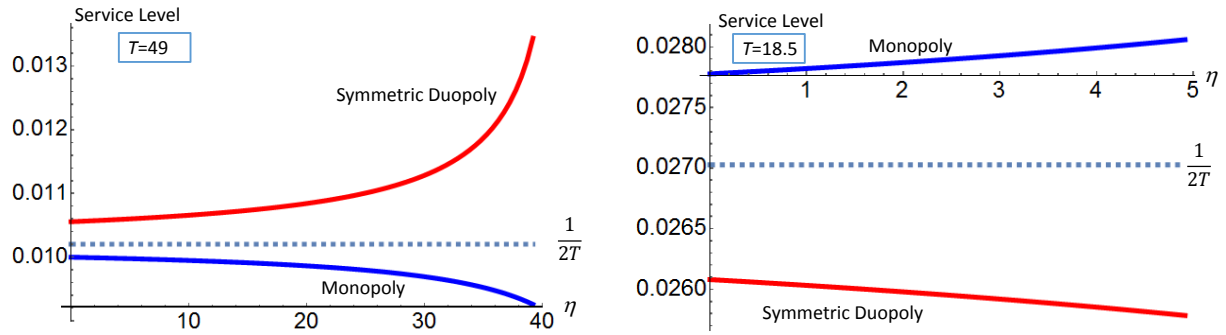

Figure 3: The Nash equilibrium service levels in a monopoly and symmetric duopoly as functions of the incentive fee parameter η ($r=200$, $b=0.9998$, $\alpha=10000$, $\rho=50000$, $T=49$ in (a)) and ($r=200$, $b=800$, $\alpha=0.5$, $\rho=3$, $T=18.5$ in (b)).

influence of incentives, as shown in Figure 3. For example, Figure 3a uses problem parameters from Figure 1 and $T=49$. In this case, the monopoly service level in the absence of incentives, θ^{M}, is less than $\frac{1}{2T}$. As a consequence, the monopoly service level decreases with incentives. On the other hand, the service level in the duopoly without incentives, $\hat{\theta}$, is greater than $\frac{1}{2T}$. As a result, the stronger the incentives, the higher the duopoly service level. In such cases, the effect of incentives in a monopoly setting is not indicative of a potential effect of the same incentive contract in the presence of competition. Another example of this phenomenon, illustrated in Figure 3b, uses the problem parameters from Figure 2. In this example, the monopoly service level in the absence of incentives is greater than $\frac{1}{2T}$. As a consequence, the monopoly service level increases as incentives strengthen. The service level in the duopoly without incentives, on the other hand, is less than $\frac{1}{2T}$, and the stronger the incentives, the lower the duopoly service level. In such cases, evaluating the impact of a performance-based contract while ignoring the competitive nature of the healthcare market may lead to the selection of contract parameters that will increase rather than decrease patient waiting times when applied in a competitive setting.

Motivated by the observations on the effects of incentives and competition from the above two examples, we formally summarize these properties in the following corollary and proposition.

COROLLARY 1. *Consider a symmetric duopoly and suppose that (62) and (66)–(69) hold. For a fixed T, let $\theta^{\mathrm{M}}(0)$ be the optimal service level for the monopoly hospital and let $(\theta^{\mathrm{NE}}(0),\theta^{\mathrm{NE}}(0))$ be the unique and symmetric Nash equilibrium for the symmetric duopoly when $\eta=0$. Then,*

(a) $\theta^{\mathrm{M}}(\eta,T)$ is increasing in η if and only if $T\geq\frac{1}{2\theta^{\mathrm{M}}(0)}$,

(b) $\theta^{\mathrm{NE}}(\eta,T)$ is increasing in η if and only if $T\geq\frac{1}{2\theta^{\mathrm{NE}}(0)}$.

Note that neither $\theta^{\mathrm{M}}(0)$ nor $\theta^{\mathrm{NE}}(0)$ depend on T. The monotonicity results in Corollary 1 have important practical implications: They show that the introduction of incentives may improve

patient services only when the waiting time target associated with the incentive policy is sufficiently weak compared to the waiting time already achieved in the absence of incentives, i.e. if T is higher than or equal to $\frac{1}{2\theta^{\mathrm{M}}(0)}$ ($\frac{1}{2\theta^{\mathrm{NE}}(0)}$). Otherwise, introducing incentives may have the unintended consequence of lowering patient service levels, with a more detrimental effect for stronger incentives.

PROPOSITION 8. *Consider a symmetric duopoly and suppose that (62) and (66)–(69) hold. For any fixed waiting time target T and incentive parameter η, let $\theta^{\mathrm{NE}}(\eta,T)$ and $\theta^{\mathrm{M}}(\eta,T)$ be the optimal Nash equilibrium service levels in the symmetric duopoly and monopoly settings, respectively.*

(a) If $\alpha \geq 1+\rho\bar{\theta}$, then competition increases service levels: $\theta^{\mathrm{NE}}(\eta,T) \geq \theta^{\mathrm{M}}(\eta,T)$.

(b) If $\alpha \leq 1$, then competition decreases service levels: $\theta^{\mathrm{NE}}(\eta,T) \leq \theta^{\mathrm{M}}(\eta,T)$.

The sufficient conditions of Proposition 8 are much simpler than those outlined in Proposition 4 as they rely only on demand parameters to characterize the effect of competition on service levels: When the demand-sensitivity parameter α is sufficiently large, competition increases service levels, and when α is sufficiently small, competition is detrimental to patient service.

In summary, when a fixed-incentive scheme defined by the waiting time target T and incentive parameter η is applied to a competitive setting, patient access to care will improve if T and α are sufficiently large but may also deteriorate if T and α are sufficiently small.

4.2. Optimal Performance-based Contract in a Symmetric Duopoly

We re-express the payer's problem in a duopoly setting as

$$\min_{T,\eta} \ \Pi(T,\eta) \equiv \sum_{i=1}^{2} \lambda_i^{(2)}\left(\boldsymbol{\theta}^{\mathrm{NE}}(T,\eta)\right) R_i\left(\boldsymbol{\theta}^{\mathrm{NE}}(T,\eta),T,\eta\right) \tag{99}$$

$$\text{s.t.} \ \ \theta_i^{\mathrm{NE}}(T,\eta) \geq \frac{1}{T}, i=1,2, \tag{100}$$

$$\pi_i^{(2)}\left(\boldsymbol{\theta}^{\mathrm{NE}}(T,\eta)\right) \geq 0, i=1,2, \tag{101}$$

$$T_l \leq T \leq T_h, \tag{102}$$

$$0 \leq \eta \leq \eta^{\max}(T), \tag{103}$$

where $\boldsymbol{\theta}^{\mathrm{NE}}(T,\eta)$ is the vector of Nash equilibrium service levels for the duopoly when the waiting time target is T and the incentive parameter is η and

$$\eta^{\max}(T) = \min\left(T, \tfrac{b}{r}\left(\alpha+1-\rho\bar{\theta}\right) - \tfrac{\rho}{2\left(\alpha+1-\rho\bar{\theta}\right)}, \tfrac{b}{r}\left(\alpha+1-3\rho\bar{\theta}\right)\right). \tag{104}$$

The first, second, and third terms on the right-hand side of (104) are obtained from (67), (68) (by letting $T \to +\infty$), and (69), respectively. This bound on the value of η, together with (62) and (66), ensures the existence and uniqueness of the Nash equilibrium service levels. Below, we

consider (99)–(103) in the monopoly and symmetric duopoly settings. Note that in the symmetric duopoly setting, the Nash equilibrium service level $\theta^{\mathrm{NE}}(\eta, T)$ is given by (91). The following result describes the optimal performance-based contract in the symmetric duopoly setting.

PROPOSITION 9. *Consider the payer's problem in the symmetric duopoly setting. Assume that (62) and (66)–(69) hold and define*

$$\eta^{\min}(T) = \frac{2b}{r}\frac{\left(\alpha + 1 - \frac{\rho}{T}\right)^2}{\left(\alpha - \frac{\rho}{T}\right)} - T \tag{105}$$

and

$$\mathcal{T} = \left(T | T \in [T_l, T_h], \eta^{\min}(T) \leq \eta^{\max}(T)\right). \tag{106}$$

(a) If $\mathcal{T} = \varnothing$ or if $T_h < \frac{1}{2\theta^{\mathrm{NE}}(0)}$, then the payer's problem has no feasible solution.

(b) If $\mathcal{T} \neq \varnothing$ and $T_l > \frac{1}{\theta^{\mathrm{NE}}(0)}$, the optimal solution for the payer's problem is

$$T^{\mathrm{opt}} \in [T_l, T_h], \eta^{\mathrm{opt}} = 0 \tag{107}$$

and the optimal value of the payer's cost is

$$\Pi^{\mathrm{opt}} = \Pi(T^{\mathrm{opt}}, \eta^{\mathrm{opt}}) = 2r\left(\alpha\theta^{\mathrm{NE}}(0) - \rho\left(\theta^{\mathrm{NE}}(0)\right)^2\right). \tag{108}$$

(c) If $\mathcal{T} \neq \varnothing$, $T_l \leq \frac{1}{\theta^{\mathrm{NE}}(0)}$ and $T_h \geq \frac{1}{2\theta^{\mathrm{NE}}(0)}$, then the optimal solution for the payer's problem is

$$T^{\mathrm{opt}} = T^*, \eta^{\mathrm{opt}} = \eta^{\min}(T^*), \tag{109}$$

where

$$T^* = \max\left(T | T \in \mathcal{T}\right) \tag{110}$$

and the optimal value of the payer's cost is

$$\Pi^{\mathrm{opt}} = \Pi(T^{\mathrm{opt}}, \eta^{\mathrm{opt}}) = 2r\left(\frac{\alpha}{T^*} - \frac{\rho}{(T^*)^2}\right). \tag{111}$$

Proposition 9 indicates that aggressive waiting time targets are incompatible with the incentive structure given by (61), while "loose" waiting time targets do not require performance incentives. On the other hand, when the payer has substantial flexibility in choosing the waiting time target (i.e. T_l is not too high and T_h is not too low), it will select the highest available waiting time T^*. The rationale for the payer to choose the highest waiting time target is that it minimizes the difference between the service level achieved in the absence of incentives and the service-level target, resulting in the lowest cost for the payer. Note that if $T_l \geq \sqrt{\frac{2b\rho}{r}}$ and $\mathcal{T} \neq \varnothing$, then $T^* = T_h$, since $\eta^{\min}(T)$ is monotone decreasing in T and $\eta^{\max}(T)$ is monotone increasing in T.

One important special case for Proposition 9 is when the waiting time target is set by the payer using medical rather than financial considerations, such as the 18-week access target used by NHS. In this case, $T_l = T_h = T$. The impact of the demand-sensitivity parameters α and ρ can be summarized as follows.

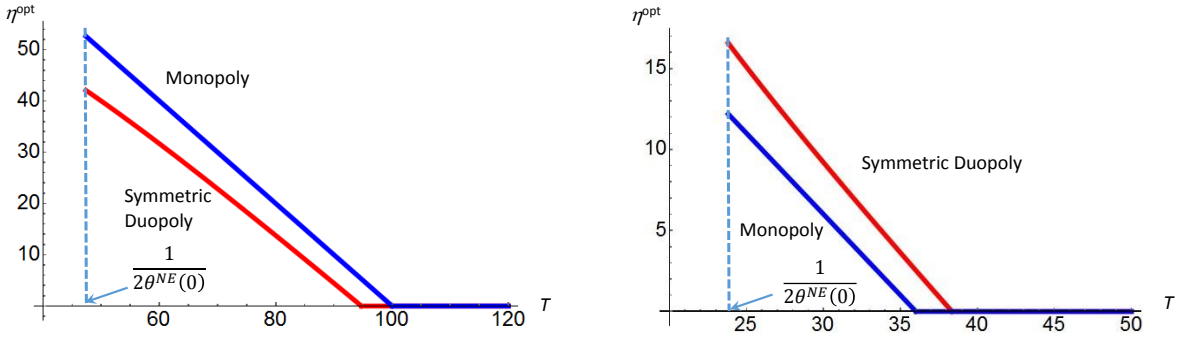

Figure 4 : The optimal incentives in a monopoly and symmetric duopoly as functions of the waiting time target T ($r = 200$, $b = 0.9998$, $\alpha = 10000$, $\rho = 50000$ in (a) and $r = 200$, $b = 800$, $\alpha = 0.5$, $\rho = 3$ in (b)).

COROLLARY 2. *Consider the payer's problem in the symmetric duopoly setting with $T_l = T_h = T$ and assume that (62) and (66)–(69) hold. For $\frac{1}{2T} \le \theta^{\mathrm{NE}}(0) \le \frac{1}{T}$ and $\eta^{\min}(T) \le \eta^{\max}(T)$, the optimal value for the incentive parameter η is increasing in α and decreasing in ρ if and only if $\alpha \ge \frac{\rho}{T} + 1$.*

Corollary 2 demonstrates that the effect of the competitive pressure parameter ρ on the optimal strength of incentives required to ensure a certain level of performance depends on patients' sensitivity to the level of access to care, α. For access-sensitive patients ($\alpha \ge 1$), the optimal incentive strength initially decreases as a new provider enters the market but eventually begins to increase as competitive pressure grows. On the other hand, if patients are not overly sensitive to the level of access to care ($\alpha < 1$), the market entrance of a competitor results in an increase in the required strength of the performance-based incentives. Figure 4 illustrates the dependence of the optimal value of the incentive parameter η as a function of the waiting time target $T = T_l = T_h$ for the two different parameter settings used in Figure 3. In both settings, higher service-level target values require higher incentive levels, but the effect of competition on the strength of the required incentives depends on the degree of patient sensitivity to the level of access to care: In the case of access-sensitive patients (Figure 4(a)) the level of incentives required is lower in a symmetric duopoly, while in the case of patients with low access sensitivity (Figure 4(b)) the presence of competition requires stronger incentives.

Figure 5 compares the optimal cost values for the payer in the cases of a monopoly and symmetric duopoly. For the monopoly case, we use the set of problem parameters that describes the access-sensitive setting in Figure 4(a), with the exception that ρ is set to 0, and calculate the optimal payer's cost value associated with two monopolistic hospitals, $\Pi^{\mathrm{M}} = 2\Pi(\eta^{\mathrm{opt}}, T^{\mathrm{opt}})$, for each set of parameter values. For the symmetric duopoly case, we calculate the optimal payer's cost value associated with a symmetric duopoly, $\Pi^{\mathrm{D}} = \Pi(\eta^{\mathrm{opt}}, T^{\mathrm{opt}})$, for the corresponding set of parameter values (with ρ now set at 50000). In other words, the ratio of $\Pi^{\mathrm{D}}/\Pi^{\mathrm{M}}$ represents the reduction in

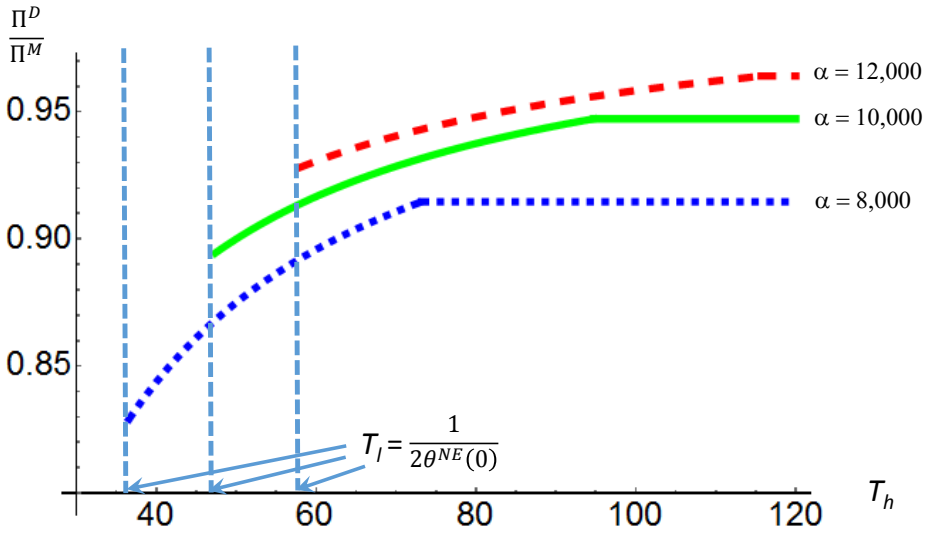

Figure 5 : The ratio of optimal payer cost values for a monopoly and symmetric duopoly as a function of the highest allowed waiting time target value T_h ($r = 200$, $b = 0.9998$, $\alpha = 10000$, $\rho = 50000$).

the payer's cost introduced by competition between two identical hospitals, with the strength of competition described by $\rho = 50000$. Figure 5 plots the ratio of $\Pi^{\mathrm{D}}/\Pi^{\mathrm{M}}$ as a function of the highest allowed waiting time target value T_h for different levels of patient access sensitivity parameter α. As Figure 5 indicates, the relative cost savings that the payer realizes due to the presence of competition are lower for more access-sensitive patients and further diminish as the waiting time target becomes weaker.

5. Discussion

The US healthcare system is experimenting with a growing number of performance-based incentive programs, which are being tested on providers competing for patients in an environment of increasing transparency about care quality. This combination of performance incentives and competition is likely to remain a dominant factor shaping healthcare delivery as payers expand the use of performance targets and, at the same time, patients, faced with increasing insurance deductibles and the increasing availability of data on provider performance, become increasingly discerning consumers of healthcare services. In our work, we examine how incentives set by a payer focused on achieving adequate patient access to care interact with the competitiveness of healthcare delivery settings.

Our analysis is based on a simple approach that models patient service dynamics at each competing provider facility, as in an $M/M/1$ queue. Patient demand for care delivered by a particular provider increases with the level of access to care the provider ensures and decreases with the level of access to care at competing facilities. It is intuitive to expect that competitive and incentive

pressures work hand in hand to improve patient access to care. Our analysis demonstrates, however, that there may be settings where this intuition does not hold. In our analysis, we look at two approaches to applying performance-based incentives to competing providers. Under the first, "soft" approach, providers are encouraged but not required to reach an access-to-care target. One example of such an approach is the contract based on waiting time targets used by the NHS in the UK (Department of Health 2008). Under the second, "strict" approach, the access-to-care target must be reached by all providers and the payer selects performance incentives to minimize the cost of achieving this target.

Using a symmetric duopoly as an example of a competitive setting, we show that under the "soft" approach, incentives and competition do indeed enhance the effect of the other in improving patient service levels when a moderate service-level target is used in an environment where patients are sensitive to the level of access to care. However, under the same approach, when an aggressive service-level target is applied in an environment where patients' care provider choices are not strongly affected by how long they must wait before accessing care, both incentives and competitive pressure may push down patient service levels. In "mixed" settings, where moderate incentives are applied to providers serving access-insensitive patients or aggressive incentives are used for providers facing access-sensitive patients, competition and incentives may move service levels in opposite directions. In the former case, in particular, the potential gain in patient service levels that incentives achieve when applied to a monopolistic provider are diminished and even reversed if used in a competitive setting.

Under the "strict" approach, all providers must reach the same access-to-care target irrespective of the degree of competition. Here, competition affects the strength of incentives and the spending required to achieve the desired performance outcomes. In particular, in the case of access-sensitive patients, competition reduces the level of incentives required, while in the case of patients with low access sensitivity, competition leads to the need for stronger incentives to achieve the same level of patient access to care. At the same time, competition benefits the payer by reducing the cost of achieving the desired level of access for either type of patient; however, the cost reduction is more pronounced in patients with low access sensitivity.

While our quantitative results rely on specific models of the demand and supply processes in healthcare settings, our qualitative conclusions, which provide detailed analysis of competitive factors in designing and implementing incentive-based programs for healthcare service providers, are based on fairly general assumptions that are likely to be broadly applicable. In particular, since many incentive-based programs are more likely to employ a "soft" approach to provider compensation than strict target enforcement, the careful matching of performance targets with

patient population type could be an important factor in generating improvements in patient access to care.

Acknowledgments

The research of the second author is partly supported by Hong Kong Research Grant Council (RGC) grant T32-102/14-N.

References

Akan, M., B. Ata, M. Lariviere. 2011. Asymmetric information and economies of scale in service contracting. *Manufacturing and Service Operations Management* **13**(1) 58–72.

Allon, G., A. Federgruen. 2007. Competition in service industries. *Operations Research* **55**(1) 37–55.

Allon, G., A. Federgruen. 2008. Service competition with general queueing facilities. *Operations Research* **56**(4) 827–849.

Allon, G., A. Federgruen. 2009. Competition in service industries with segmented markets. *Management Science* **55**(4) 619–634.

Anderson, G.F., U.E. Reinhardt, P.S. Hussey, V. Petrosyan. 2003. It's the prices, stupid: Why the United States is so different from other countries. *Health Affairs* **22**(3) 89–105.

Andritsos, D.A., S. Aflaki. 2015. Competition and the operational performance of hospitals: The role of hospital objectives. *Production and Operations Management* **24**(11) 1812–1832.

Bernstein, F., A. Federgruen. 2004. A general equilibrium model for industries with price and service competition. *Operations Research* **52**(6) 868–886.

Berwick, D.M. 2011. Launching accountable care organizations – The proposed rule for the Medicare Shared Savings Program. *New England Journal of Medicine* 364:e32

Bloom, N., Z. Cooper, M. Gaynor, S. Gibbons, S. Jones, A. McGuire, R. Moreno-Serra, C. Propper, J. Van Reenen, S. Seiler. 2011. In defence of our research on competition in England's National Health Service: Author's reply. *Lancet* **378** 2064–2065.

Brekke, K., L. Siciliani, O.R. Straume. 2008. Competition and waiting times in hospital markets. *Journal of Public Economics* **92**(7) 1607–1628.

Bynum, J.P.W., E. Meara, C.-H. Chang, J.M. Rhoads. 2016. *Our Parents, Ourselves: Health Care for an Aging Population.* February 17, 2016. A Report of the Dartmouth Atlas Project. The Dartmouth Institute for Health Policy and Clinical Practice. `http://www.dartmouthatlas.org/downloads/reports/Our_Parents_Ourselves_021716.pdf`. Accessed April 2, 2016.

Cachon, G., P. Harker. 2002. Competition and outsourcing with scale economies, *Management Science*, **48**(10) 1314-1333.

Cachon, G., S. Netessine. 2004. Game theory in supply chain analysis. In: Simchi-Levi, D., S.D. Wu, and M. Shen (eds.), *Handbook of Quantitative Supply Chain Analysis: Modeling in the E-Business Era.* Kluwer.

Cashin, C., Y-L. Chi, P.C. Smith, M. Borowitz, S. Thomson (eds). 2014. *Paying for Performance in Health Care: Implications for Health System Performance and Accountability*. European Observatory on health Systems and Policies Series (OECD-WHO), Open University Press, United Kingdom.

Chalkley, M., J.M. Malcomson. 1998. Contracting for health services with unmonitored quality. *The Economic Journal*, **108**(449) 1093–1110.

Centers for Medicare and Medicaid Services. 2016. What is Hospital Compare? `https://www.medicare.gov/hospitalcompare/About/What-Is-HOS.html`. Accessed April 2, 2016.

Centers for Medicare and Medicaid Services. 2016. Bundled payments for care improvement (BPCI) initiative: General information. `https://innovation.cms.gov/initiatives/Bundled-Payments/index.html`. Accessed April 2, 2016.

Centers for Medicare and Medicaid Services. 2016. Hospital-value based purchasing. `https://www.cms.gov/Medicare/Quality-initiatives-patient-assessment-instruments/hospital-value-based-purchasing/index.html`. Accessed April 2, 2016.

Cooper, Z., S. Gibbons, S. Jones, A. McGuire. 2011. Does hospital competition save lives? Evidence from the English NHS patient choice reforms. *The Economic Journal* **121**(554) 228-260.

De Fraja, G. 2000. Contracts for healthcare and asymmetric information. *Journal of Health Economics* **19**(5), 663–677.

Department of Health. 2008. *Standard NHS Contract for Acurate Services (2008/2009)* `http://dh.gov.uk/en/publicationsandstatistics/publications/publicationspolicyandguidance/dh_091451`. Accessed April 2, 2016.

Fisher, E.S., D.E. Wennberg, T.A. Stukel, D.J. Gottlieb, F.L. Lucas, E.L. Pinder. 2003. The implications of regional variations in Medicare spending. Part 1. *Annanls Internal Medicine* **138**:273–287.

Fuloria, P. C., S. A. Zenios. 2001. Outcomes-adjusted reimbursement in a health-care delivery system. *Management Science* **47**(6), 735–751.

Gaynor, M., R. Moreno-Serra, C. Propper. 2013. Death by market power: Reform, competition and patient outcomes in the National Health Service. *American Economic Journal: Economic Policy* **5**(4) 134–166.

Gaynor M., R. J. Town. 2012. Competition in health care markets. In McGuire T., M. Pauly, and P.P. Barros (eds), *Handbook of Health Economics,*. Vol 2. Elsevier/North-Holland: Amsterdam.

Hasija, S., E.J. Pinker, R. A. Shumsky. 2008. Call center outsourcing contracts under asymmetric information. *Management Science* **54**(4) 793–807.

H.R. 3590, 111th Cong.: Patient Protection and Affordable Care Act. 2010. `http://www.govtrack.us/congress/bill.xpd?bill=h111-3590&tab=reports`. Accessed April 2, 2016.

Jiang, H., Z. Pang, S. Savin. 2012. Performance-based contracts for outpatient medical services. *Manufacturing and Service Operations Management* **14**(2) 654–669.

Kolstad, C., L. Mathiesen. 1987. Necessary and sufficient conditions for uniqueness in a Cournot equilibrium. *Review of Economic Studies* **54** 681–690.

Laffont, J.-J., J. Tirole. 1993. *A Theory of Incentives in Procurement and Regulation.* MIT Press, Cambridge, MA.

Lee, T., A.E. Lechner, E.R. Boukus. 2013. *The Surge in Urgent Care Centers: Emergency Department Alternative or Costly Convenience?* Center for Study of Health System Change, Research Brief 26. `http://www.hschange.com/CONTENT/1366/1366.pdf` (retrieved on April 2, 2016).

Lee, D. K. K., S. A. Zenios. 2012. An evidence-based incentive system for Medicare's end-stage renal disease program. *Management Science* **58**(6), 1092–1105.

Lewis, R., J. Appleby. 2006. Can the English NHS meet the 18-week waiting list target? *Journal of the Royal Society of Medicine* **99** 10-13.

Meritt Hawkins. 2014. *Physician Appointment Wait Times and Medicaid and Medicare Acceptance Rates.* `http://www.merritthawkins.com/uploadedFiles/MerrittHawkings/Surveys/mha2014waitsurvPDF.pdf`. Accessed April 2, 2016.

Miraldo, M., L. Siciliani, A. Street. 2011. Price adjustment in the hospital sector. *Journal of Health Economics* **30** 112–125.

National Health Service. 2016. *Your Choices in the NHS.* `http://www.nhs.uk/choiceintheNHS/Yourchoices/Pages/your-choices.aspx`. Accessed April 2, 2016.

OECD. 2015. *Health at a Glance 2015: OECD Indicators, OECD Publishing*, Paris, p. 164. `http://dx.doi.org/10.1787/health_glance-2015-en`. Accessed April 2, 2016.

O'Reilly K.B. 2010. Posting emergency wait times: good marketing or good medicine? *American Medical News.* Amednews.com. `http://www.ama-assn.org/amednews/2010/10/11/prl21011.htm`. Accessed April 2, 2016.

Price Transparency in Health Care. 2014. *Report from the HFMA Price Transparency Task Force.* `https://www.hfma.org/transparency/` (retrieved on April 2, 2016).

Propper, C., S. Burgess, D. Gossage. 2008. Competition and quality: Evidence from the NHS internal market 1991–9*. *The Economic Journal* **118**(525) 138–170.

Propper C., S. Burgess, K. Green. 2004. Does competition between hospitals improve the quality of care? Hospital death rates and the NHS internal market. *Journal of Public Economics* 88 1247–1272.

Propper, C., M. Sutton, C. Whitnall, F. Windmeijer. 2008. Did "targets and terror" reduce waiting times in England for hospital care? *The B.E. Journal of Economic Analysis and Policy* **8**(2) 1935-1682.

Propper C., M. Sutton, C. Whitnall, F. Windmeijer. 2010. Incentives and targets in hospital care: Evidence from a natural experiment. *Journal of Public Economics* **94** 318–335.

Ren, Z.J., Y.-P. Zhou. 2008. Call center outsourcing: Coordinating staffing level and service quality. *Management Science* **54**(2) 369–383.

Schenker, Y., R.M. Arnold, A.J. London. 2014. The ethics of advertising for healthcare services. *American Journal of Bioethics*. **14** (3) 34–43.

So, K.C., C. S. Tang. 2000. Modeling the impact of an outcome-oriented reimbursement policy on clinic, patients, and pharmaceutical firms. *Management Science* **46**(7) 875–892.

Thomson, S., R. Osborn, D. Squires, M. Jun. 2013. *International Profile of Health Care Systems*, The Commonwealth Fund.

Topkis, D. 1998. *Supermodularity and Complementarity*. Princeton University Press, Princeton, NJ.

Vaida, B., A. Weiss. 2015. Health Care Consolidation. `http://www.allhealth.org/publications/Consolidation-Toolkit_169.pdf`. Accessed April 2, 2016.

Vives, X. 2001. *Oligopoly Pricing: Old Ideas and New Tools*, p. 150. MIT Press, Cambridge, MA.

Wennberg, J.E.. 2002. Unwarranted variations in healthcare delivery: implications for academic medical centres. *British Med. Journal* **352**:961–964.

Wennberg, J.E., E.S. Fisher, S.M. Sharp, T.A. Bubolz, V.J. Fusca III, D.C. Goodman, D.J. Gottlieb, J. Lan, Z. Peng, S.R. Raymond, J.S. Skinner, T.S.D. Wang, P. Wright-Slaughter. 2006. *The Care of Patients with Severe Chronic Illness: An Online Report on the Medicare Program by the Dartmouth Atlas Project*. The Center for the Evaluative Clinical Sciences. `http://www.dartmouthatlas.org/downloads/atlases/2006_Chronic_Care_Atlas.pdf`. Accessed April 2, 2016.

Appendix: Proofs

Proof of Proposition 1

In the proof below, we omit the dependence of fee and profit functions on T and $\boldsymbol{\xi}$ to simplify the notation. As follows from Theorem 1 and Theorem 5, respectively, in Cachon and Netessine (2004), a Nash equilibrium exists if $\pi_i^{(n)}(\theta)$ is concave in θ_i for any hospital i, and such an equilibrium is unique if the Hessian matrix of $\pi_i^{(n)}(\theta)$ is diagonally strictly dominant. Thus, the sufficient condition for the existence and uniqueness of the Nash equilibrium is

$$\frac{\partial^2 \pi_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i^2} + \sum_{j \neq i} \left| \frac{\partial^2 \pi_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right| < 0, \forall i, j \neq i, \theta \in \boldsymbol{\Theta}. \tag{A1}$$

We proceed to derive the conditions that ensure that (A1) holds. For the expression on the left-hand side of (A1), we get

$$\begin{aligned}
&\frac{\partial^2 \pi_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i^2} + \sum_{j \neq i} \left| \frac{\partial^2 \pi_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right| \\
&= \frac{\partial^2 \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i^2} [R_i(\theta_i) - C_i'(\lambda_i^{(n)}(\boldsymbol{\theta}) + \theta_i)] - C_i''(\lambda_i^{(n)}(\boldsymbol{\theta}) + \theta_i) \left(\frac{\partial \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i} + 1 \right)^2 \\
&\qquad + \lambda_i^{(n)}(\boldsymbol{\theta}) R_i''(\theta_i) + 2 \frac{\partial \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i} R_i'(\theta_i) \\
&+ \sum_{j \neq i} \left\{ \left| \frac{\partial^2 \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} [R_i(\theta_i) - C_i'(\lambda_i^{(n)}(\boldsymbol{\theta}) + \theta_i)] + \frac{\partial \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_j} R_i'(\theta_i) \right. \right. \\
&\qquad \left. \left. - C_i''(\lambda_i^{(n)}(\boldsymbol{\theta}) + \theta_i) \left[1 + \frac{\partial \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i} \right] \frac{\partial \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_j} \right| \right\} \\
&\leq \frac{\partial^2 \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i^2} [R_i(\theta_i) - C_i'(\lambda_i^{(n)}(\boldsymbol{\theta}) + \theta_i)] - C_i''(\lambda_i^{(n)}(\boldsymbol{\theta}) + \theta_i) \left(\frac{\partial \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i} + 1 \right)^2 \\
&\qquad + \lambda_i^{(n)}(\boldsymbol{\theta}) R_i''(\theta_i) + 2 \frac{\partial \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i} R_i'(\theta_i) \\
&+ \sum_{j \neq i} \left\{ \left| \frac{\partial^2 \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right| [R_i(\theta_i) - C_i'(\lambda_i^{(n)}(\boldsymbol{\theta}) + \theta_i)] - \frac{\partial \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_j} R_i'(\theta_i) \right. \\
&\qquad \left. - C_i''(\lambda_i^{(n)}(\boldsymbol{\theta}) + \theta_i) \left[1 + \frac{\partial \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i} \right] \frac{\partial \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_j} \right\} \\
&= \left[\frac{\partial^2 \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i^2} - \sum_{j \neq i} \frac{\partial^2 \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right] [R_i(\theta_i) - C_i'(\lambda_i^{(n)}(\boldsymbol{\theta}) + \theta_i)] + \left[2 \frac{\partial \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i} - \sum_{j \neq i} \frac{\partial \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_j} \right] R_i'(\theta_i) \\
&\quad - C_i''(\lambda_i^{(n)}(\boldsymbol{\theta}) + \theta_i) \left[1 + \frac{\partial \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i} \right] \left[1 + \sum_k \frac{\partial \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_k} \right] + \lambda_i^{(n)}(\boldsymbol{\theta}) R_i''(\theta_i) \\
&\leq \left[\frac{\partial^2 \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i^2} - \sum_{j \neq i} \frac{\partial^2 \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right] [R_i(\theta_i) - C_i'(\lambda_i^{(n)}(\boldsymbol{\theta}) + \theta_i)] + \left[2 \frac{\partial \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_i} - \sum_{j \neq i} \frac{\partial \lambda_i^{(n)}(\boldsymbol{\theta})}{\partial \theta_j} \right] R_i'(\theta_i)
\end{aligned}$$

$$
\begin{aligned}
&-C_i''(\lambda_i^{(n)}(\boldsymbol{\theta})+\theta_i)\left[1+\frac{\partial\lambda_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_i}\right]\left[1+\sum_k\frac{\partial\lambda_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_k}\right]\\
\le&\left[\frac{\partial^2\lambda_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_i^2}-\sum_{j\neq i}\frac{\partial^2\lambda_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right][R_i(\theta_i)-C_i'(\lambda_i^{(n)}(\boldsymbol{\theta})+\theta_i)]+3\frac{\partial\lambda_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_i}R_i'(\theta_i)\\
&-C_i''(\lambda_i^{(n)}(\boldsymbol{\theta})+\theta_i)\left[1+\frac{\partial\lambda_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_i}\right]\left[1+\sum_k\frac{\partial\lambda_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_k}\right]\\
\le&\left[\frac{\partial^2\lambda_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_i^2}-\sum_{j\neq i}\frac{\partial^2\lambda_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right][R_i(\theta_i)-C_i'(\lambda_i^{(n)}(\boldsymbol{\theta})+\theta_i)]-\frac{\partial\lambda_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_i}[C_i''(\lambda_i^{(n)}(\boldsymbol{\theta})+\theta_i)-3R_i'(\theta_i)]\\
\le&\left[-\delta_{ii}+\sum_{j\neq i}\delta_{ij}\right]\left[R_i(\theta_i)-C_i'\left(\lambda_i^{(n)}(\boldsymbol{\theta})+\theta_i\right)\right]-\left[d_i-3\bar{R}_i'\right]\left[\frac{\partial\lambda_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_i}\right],
\end{aligned}
\tag{A2}
$$

where the second inequality follows from the concavity of R_i, the third, from Assumption 4, the fourth, from Assumption 4, which implies $1+\sum_k\frac{\partial\lambda_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_k}\ge 1$, and the last, from (12), (13), (18) and (31). The right-hand side of (A2) is negative if $\delta_{ii}\ge\sum_{j\neq i}\delta_{ij}$ and $d_i>3\bar{R}_i'$.

On the other hand, if $\delta_{ii}<\sum_{j\neq i}\delta_{ij}$ and $d_i>3\bar{R}_i'$, we have

$$
\begin{aligned}
&\left[-\delta_{ii}+\sum_{j\neq i}\delta_{ij}\right]\left[R_i(\theta_i)-C_i'\left(\lambda_i^{(n)}(\boldsymbol{\theta})+\theta_i\right)\right]-\left[d_i-3\bar{R}_i'\right]\left[\frac{\partial\lambda_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_i}\right]\\
\le&\left[-\delta_{ii}+\sum_{j\neq i}\delta_{ij}\right][\bar{R}_i-c_i]-\gamma_{ii}[d_i-3\bar{R}_i'],
\end{aligned}
\tag{A3}
$$

Note that

$$
\left[-\delta_{ii}+\sum_{j\neq i}\delta_{ij}\right][\bar{R}_i-c_i]-\gamma_{ii}[d_i-3\bar{R}_i']<0 \tag{A4}
$$

if and only if

$$
\gamma_{ii}[d_i-3\bar{R}_i']>\left[-\delta_{ii}+\sum_{j\neq i}\delta_{ij}\right][\bar{R}_i-c_i], \tag{A5}
$$

or

$$
d_i-3\bar{R}_i'>\frac{\left[-\delta_{ii}+\sum_{j\neq i}\delta_{ij}\right][\bar{R}_i-c_i]}{\gamma_{ii}}. \tag{A6}
$$

Note that this condition is a stronger version of $d_i-3\bar{R}_i'>0$. (A6) is equivalent to

$$
3\bar{R}_i'<d_i-\frac{\left[-\delta_{ii}+\sum_{j\neq i}\delta_{ij}\right][\bar{R}_i-c_i]}{\gamma_{ii}}, \tag{A7}
$$

or

$$
\bar{R}_i'<\frac{d_i}{3}-\frac{\left[-\delta_{ii}+\sum_{j\neq i}\delta_{ij}\right][\bar{R}_i-c_i]}{3\gamma_{ii}}. \tag{A8}
$$

Using $(-\delta_{ii}+\sum_{j\neq i}\delta_{ij})^+ = \max(-\delta_{ii}+\sum_{j\neq i}\delta_{ij},0)$, we can combine the results for $-\delta_{ii}+\sum_{j\neq i}\delta_{ij}\leq 0$ and $-\delta_{ii}+\sum_{j\neq i}\delta_{ij}>0$ as

$$\bar{R}_i' < \frac{d_i}{3} - \frac{\left[-\delta_{ii}+\sum_{j\neq i}\delta_{ij}\right]^+ [\bar{R}_i - c_i]}{3\gamma_{ii}}. \tag{A9}$$

□

Proof of Proposition 2

First, note that (38)–(40) ensure the existence and uniqueness of the symmetric Nash equilibrium in settings both with and without performance-based incentives. Because we analyze a symmetric setting, in the following we drop the index i for the functions associated with a particular hospital. We also omit the dependence on T and ξ. To facilitate the analysis, define $\Psi_i^{(n)}(t)$ as the first-order derivative of $\pi_i^{(n)}$:

$$\Psi_i^{(n)}(t) = \frac{\partial\lambda^{(n)}(t,...,t)}{\partial\theta_i}\left[R(t) - C'(\lambda^{(n)}(t,...,t)+t)\right] - C'(\lambda^{(n)}(t,...,t)+t) + \lambda^{(n)}(t,...,t)R'(t). \tag{A10}$$

The concavity of the objective function and symmetry imply that if the equilibrium service level is an interior point of the interval $\left[0,\bar{\theta}\right]$, it must satisfy the first-order condition:

$$\Psi_i^{(n)}(t) = 0. \tag{A11}$$

If $\Psi_i^{(n)}(\theta^{(n)})<0$, then the concavity of $\pi_i^{(n)}$ in θ_i implies that $\theta^{(n)}=0$; and if $\Psi_i^{(n)}(\theta^{(n)})>0$, then the concavity of $\pi_i^{(n)}$ in θ_i implies that $\theta^{(n)}=\bar{\theta}$.

Next, we show that $\Psi_i^{(n)}(t)$ is a strictly decreasing function of t.

$$\begin{aligned}
\frac{d\Psi_i^{(n)}(t)}{dt} &= \left[\frac{\partial^2\lambda^{(n)}(t,...,t)}{\partial\theta_i^2} + \sum_{j\neq i}\frac{\partial^2\lambda^{(n)}(t,...,t)}{\partial\theta_i\partial\theta_j}\right][R(t) - C'(\lambda^{(n)}(t,...,t)+t)] \\
&\quad - C''(\lambda^{(n)}(t,...,t)+t)\left[\frac{\partial\lambda^{(n)}(t,...,t)}{\partial\theta_i}+1\right]\left[\frac{\partial\lambda^{(n)}(t,...,t)}{\partial\theta_i} + \sum_{j\neq i}\frac{\partial\lambda^{(n)}(t,...,t)}{\partial\theta_j}+1\right] \\
&\quad + \lambda^{(n)}(t,...,t)R''(t) + \left[2\frac{\partial\lambda^{(n)}(t,...,t)}{\partial\theta_i} + \sum_{j\neq i}\frac{\partial\lambda^{(n)}(t,...,t)}{\partial\theta_j}\right]R'(t) \\
&\leq -(\delta+(n-1)\bar{\Delta})\left[\underline{R}-\bar{c}\right] - d\left[\gamma+1\right] + (2\bar{\gamma}-(n-1)\bar{\Gamma})\bar{R}',
\end{aligned} \tag{A12}$$

where the inequality is due to Assumption 4 and the fact that $\frac{\partial^2\lambda^{(n)}(t,...,t)}{\partial\theta_i^2}\leq -\delta$, $\frac{\partial^2\lambda^{(n)}(t,...,t)}{\partial\theta_i\partial\theta_j}\leq -\bar{\Delta}$, $R(t)-C'(\lambda^{(n)}(t,...,t)+t)\geq \underline{R}-\bar{c}>0$, $C''(\lambda^{(n)}(t,...,t)+t)\geq d$, $\gamma\leq\frac{\partial\lambda^{(n)}(t,...,t)}{\partial\theta_i}\leq\bar{\gamma}$, R is concave in t, and $-\Gamma\leq\frac{\partial\lambda^{(n)}(t,...,t)}{\partial\theta_j}\leq-\bar{\Gamma}$. The last expression in (A12) is strictly smaller than zero if and only if

$$\bar{R}' < \frac{(\delta+(n-1)\bar{\Delta})\left[\underline{R}-\bar{c}\right] + d\left[\gamma+1\right]}{2\bar{\gamma}-(n-1)\bar{\Gamma}}.$$

The strict monotonicity of $\Psi^{(n)}(t)$ implies that if $\Psi^{(n)}(0)<0$, then $\theta^{(n)}=0$, and otherwise, $\theta^{(n)}$ can be chosen as the greatest value of t in $[0,\bar{\theta}]$ such that $\Psi^{(n)}(t)\geq 0$.

We now explore the effect of introducing performance-based incentives. The first derivative of the profit function in the absence of incentives is

$$\tilde{\Psi}_i^{(n)}(t)=\frac{\partial\lambda^{(n)}(t,...,t)}{\partial\theta_i}\left[R_{\rm b}-C'(\lambda^{(n)}(t,...,t)+t)\right]-C'(\lambda^{(n)}(t,...,t)+t). \tag{A13}$$

Similarly to the analysis for $\Psi_i^{(n)}$, we can show that $\tilde{\Psi}_i^{(n)}(t)$ is also strictly decreasing in t.

The strict monotonicity of $\Psi_i^{(n)}$ and $\tilde{\Psi}_i^{(n)}$ implies that $\theta^{(n)}\geq\theta_{\rm b}^{(n)}$ if $\Psi_i^{(n)}\left(\theta_{\rm b}^{(n)}\right)\geq\tilde{\Psi}_i^{(n)}\left(\theta_{\rm b}^{(n)}\right)$ and $\theta^{(n)}\leq\theta_{\rm b}^{(n)}$ otherwise.

Note that

$$\Psi_i^{(n)}\left(\theta_{\rm b}^{(n)}\right)-\tilde{\Psi}_i^{(n)}\left(\theta_{\rm b}^{(n)}\right)=\frac{\partial\lambda^{(n)}\left(\theta_{\rm b}^{(n)},...,\theta_{\rm b}^{(n)}\right)}{\partial\theta_i}\left[R\left(\theta_{\rm b}^{(n)}\right)-R_{\rm b}\right]+\lambda^{(n)}\left(\theta_{\rm b}^{(n)},...,\theta_{\rm b}^{(n)}\right)R'\left(\theta_{\rm b}^{(n)}\right), \tag{A14}$$

which is nonnegative if

$$R'(\theta_{\rm b}^{(n)})\geq\frac{\frac{\partial\lambda^{(n)}(\theta_{\rm b}^{(n)},...,\theta_{\rm b}^{(n)})}{\partial\theta_i}}{\lambda^{(n)}(\theta_{\rm b}^{(n)},...,\theta_{\rm b}^{(n)})}\left[R_{\rm b}-R\left(\theta_{\rm b}^{(n)}\right)\right] \tag{A15}$$

and nonpositive otherwise, which leads to the desired results. □

Proof of Proposition 3

Assumption $\eta\leq\eta^\star(T)$ ensures the existence and uniqueness of the symmetric Nash equilibrium. Similarly to the Proof of Proposition 2, we drop the index i for the hospital demand and cost functions. The concavity of the objective function and symmetry imply that if the equilibrium service rate is an interior point of the interval $\left[0,\bar{\theta}\right]$, it must satisfy the first-order condition:

$$\frac{\partial\lambda^{(n)}(t,...,t)}{\partial\theta_i}\left[r-C'(\lambda^{(n)}(t,...,t)+t)\right]-C'(\lambda^{(n)}(t,...,t)+t)+\left[\lambda^{(n)}(t,...,t)T-\frac{\partial\lambda^{(n)}(t,...,t)}{\partial\theta_i}\right]\eta re^{-tT}=0. \tag{A16}$$

Define $\Psi_i^{(n)}(t|\eta)$ as the first-order derivative of $\pi_i^{(n)}$:

$$\begin{aligned}\Psi_i^{(n)}(t|\eta)=&\frac{\partial\lambda^{(n)}(t,...,t)}{\partial\theta_i}\left[r-C'(\lambda^{(n)}(t,...,t)+t)\right]-C'(\lambda^{(n)}(t,...,t)+t)\\&+\left[\lambda^{(n)}(t,...,t)T-\frac{\partial\lambda^{(n)}(t,...,t)}{\partial\theta_i}\right]\eta re^{-tT}.\end{aligned} \tag{A17}$$

Next, we show that $\Psi_i^{(n)}(t|\eta)$ is a strictly decreasing function of t.

$$\begin{aligned}\frac{d\Psi_i^{(n)}(t|\eta)}{dt}=&\left[\frac{\partial^2\lambda^{(n)}(t,...,t)}{\partial\theta_i^2}+\sum_{j\neq i}\frac{\partial^2\lambda^{(n)}(t,...,t)}{\partial\theta_i\partial\theta_j}\right]\left[r-\eta re^{-tT}-C'(\lambda^{(n)}(t,...,t)+t)\right]\\&-C''(\lambda^{(n)}(t,...,t)+t)\left[\frac{\partial\lambda^{(n)}(t,...,t)}{\partial\theta_i}+1\right]\left[\frac{\partial\lambda^{(n)}(t,...,t)}{\partial\theta_i}+\sum_{j\neq i}\frac{\partial\lambda^{(n)}(t,...,t)}{\partial\theta_j}+1\right]\end{aligned}$$

$$
\begin{aligned}
&-\lambda^{(n)}(t,...,t)T^2\eta r e^{-tT} + \left[2\frac{\partial\lambda^{(n)}(t,...,t)}{\partial\theta_i} + \sum_{j\neq i}\frac{\partial\lambda^{(n)}(t,...,t)}{\partial\theta_j}\right] T\eta r e^{-tT} \\
&\leq -(\delta+(n-1)\bar{\Delta})\left[r-\eta r-\bar{c}\right] - d\left[\gamma+1\right] + (2\bar{\gamma}-(n-1)\bar{\Gamma})T\eta r \\
&= \left[(2\bar{\gamma}-(n-1)\bar{\Gamma})T + (\delta+(n-1)\bar{\Delta})\right]\eta r - (\delta+(n-1)\bar{\Delta})[r-\bar{c}] - d\left[\gamma+1\right], \qquad \text{(A18)}
\end{aligned}
$$

where the first inequality is due to Assumption 4 and the fact that $\frac{\partial^2\lambda^{(n)}(t,...,t)}{\partial\theta_i^2} \leq -\delta$, $\frac{\partial^2\lambda^{(n)}(t,...,t)}{\partial\theta_i\partial\theta_j} \leq -\bar{\Delta}$, $r-\eta r e^{-tT} - C'(\lambda^{(n)}(t,...,t)+t) \geq r-\eta r-\bar{c}>0$, $C''(\lambda^{(n)}(t,...,t)+t) \geq d$, $\gamma \leq \frac{\partial\lambda^{(n)}(t,...,t)}{\partial\theta_i} \leq \bar{\gamma}$, and $-\Gamma \leq \frac{\partial\lambda^{(n)}(t,...,t)}{\partial\theta_j} \leq -\bar{\Gamma}$. The last expression in (A18) is strictly smaller than zero if and only if

$$
\eta < \frac{(\delta+(n-1)\bar{\Delta})\left[1-\frac{\bar{c}}{r}\right] + \frac{d}{r}\left[\gamma+1\right]}{(2\bar{\gamma}-(n-1)\bar{\Gamma})T + (\delta+(n-1)\bar{\Delta})}. \qquad \text{(A19)}
$$

The strict monotonicity of $\Psi^{(n)}(\cdot|\eta)$ implies that $\theta^{(n)}(\eta)>0$ if and only if $\Psi^{(n)}(0|\eta)>0$ and that $\theta^{(n)}(\eta)<\bar{\theta}$ if and only if $\Psi^{(n)}\left(\bar{\theta}|\eta\right)<0$.

We now derive sufficient conditions for the monotonicity of $\theta^{(n)}(\eta)$ with respect to η. As follows from (41) and (42), $\theta^{(n)}(\eta) \geq \theta^{(n)}(0) = \theta_{\mathrm{b}}^{(n)}$ if $T \geq T^\star$ and $\theta^{(n)}(\eta) \leq \theta^{(n)}(0) = \theta_{\mathrm{b}}^{(n)}$ otherwise.

Note that $\lambda^{(n)}(t,...,t)T - \frac{\partial\lambda^{(n)}(t,...,t)}{\partial\theta_i}$ is increasing in t. Then, it follows from Assumption 3 and Assumption 4 that for $T \geq T^\star$,

$$
\begin{aligned}
&\lambda^{(n)}(\theta^{(n)}(\eta),...,\theta^{(n)}(\eta))T - \frac{\partial\lambda^{(n)}(\theta^{(n)}(\eta),...,\theta^{(n)}(\eta))}{\partial\theta_i} \\
&\geq \lambda^{(n)}(\theta^{(n)}(0),...,\theta^{(n)}(0))T - \frac{\partial\lambda^{(n)}(\theta^{(n)}(0),...,\theta^{(n)}(0))}{\partial\theta_i}, \qquad \text{(A20)}
\end{aligned}
$$

while for $T < T^\star$, we have

$$
\begin{aligned}
&\lambda^{(n)}(\theta^{(n)}(\eta),...,\theta^{(n)}(\eta))T - \frac{\partial\lambda^{(n)}(\theta^{(n)}(\eta),...,\theta^{(n)}(\eta))}{\partial\theta_i} \\
&< \lambda^{(n)}(\theta^{(n)}(0),...,\theta^{(n)}(0))T - \frac{\partial\lambda^{(n)}(\theta^{(n)}(0),...,\theta^{(n)}(0))}{\partial\theta_i}. \qquad \text{(A21)}
\end{aligned}
$$

For any $\eta > \eta'$, we have

$$
\begin{aligned}
&\Psi_i^{(n)}(\theta^{(n)}(\eta)|\eta) - \Psi_i^{(n)}(\theta^{(n)}(\eta)|\eta') \\
&= \left[\lambda^{(n)}(\theta^{(n)}(\eta),...,\theta^{(n)}(\eta))T - \frac{\partial\lambda^{(n)}(\theta^{(n)}(\eta),...,\theta^{(n)}(\eta))}{\partial\theta_i}\right](\eta-\eta')re^{-\theta^{(n)}(\eta)T}. \qquad \text{(A22)}
\end{aligned}
$$

If $T \geq T^*$, then the fact that (A20) holds shows that

$$
\begin{aligned}
&\Psi_i^{(n)}(\theta^{(n)}(\eta)|\eta) - \Psi_i^{(n)}(\theta^{(n)}(\eta)|\eta') \\
&= \left[\lambda^{(n)}(\theta^{(n)}(\eta),...,\theta^{(n)}(\eta))T - \frac{\partial\lambda^{(n)}(\theta^{(n)}(\eta),...,\theta^{(n)}(\eta))}{\partial\theta_i}\right](\eta-\eta')re^{-\theta^{(n)}(\eta)T} \\
&\geq \left[\lambda^{(n)}(\theta^{(n)}(0),...,\theta^{(n)}(0))T - \frac{\partial\lambda^{(n)}(\theta^{(n)}(0),...,\theta^{(n)}(0))}{\partial\theta_i}\right](\eta-\eta')re^{-\theta^{(n)}(\eta)T} \\
&= \lambda^{(n)}(\theta^{(n)}(0),...,\theta^{(n)}(0))(T-T^*)(\eta-\eta')re^{-\theta^{(n)}(\eta)T} \geq 0. \qquad \text{(A23)}
\end{aligned}
$$

If $T < T^*$, then the fact that inequality (A21) holds implies that

$$\begin{aligned}
&\Psi_i^{(n)}(\theta^{(n)}(\eta)|\eta) - \Psi_i^{(n)}(\theta^{(n)}(\eta)|\eta') \\
&= \left[\lambda^{(n)}(\theta^{(n)}(\eta),...,\theta^{(n)}(\eta))T - \frac{\partial\lambda^{(n)}(\theta^{(n)}(\eta),...,\theta^{(n)}(\eta))}{\partial\theta_i}\right](\eta-\eta')re^{-\theta^{(n)}(\eta)T} \\
&\leq \left[\lambda^{(n)}(\theta^{(n)}(0),...,\theta^{(n)}(0))T - \frac{\partial\lambda^{(n)}(\theta^{(n)}(0),...,\theta^{(n)}(0))}{\partial\theta_i}\right](\eta-\eta')re^{-\theta^{(n)}(\eta)T} \\
&= \lambda^{(n)}(\theta^{(n)}(0),...,\theta^{(n)}(0))(T-T^*)(\eta-\eta')re^{-\theta^{(n)}(\eta)T} \leq 0. \qquad \text{(A24)}
\end{aligned}$$

Thus, we have proved that

$$\Psi_i^{(n)}(\theta^{(n)}(\eta)|\eta) - \Psi_i^{(n)}(\theta^{(n)}(\eta)|\eta') \qquad \text{(A25)}$$

is greater than or equal to zero if $T \geq T^\star$ and less than or equal to zero otherwise. The monotonicity of $\Psi_i^{(n)}(\cdot|\eta)$ implies that $\theta^{(n)}(\eta) \geq \theta^{(n)}(\eta')$ if $T \geq T^\star$, and $\theta^{(n)}(\eta) \leq \theta^{(n)}(\eta')$ otherwise. To see this, when $T \geq T^\star$, we consider two cases. If $\theta^{(n)}(\eta) < \bar{\theta}$, the monotonicity of $\Psi_i^{(n)}(\cdot|\eta)$ implies that $\Psi_i^{(n)}(\theta^{(n)}(\eta)|\eta) \leq 0$ and then $\Psi_i^{(n)}(\theta^{(n)}(\eta)|\eta') \leq 0$. The uniqueness of the solution and monotonicity of $\Psi_i^{(n)}(\cdot|\eta')$ immediately imply that $\theta^{(n)}(\eta') \leq \theta^{(n)}(\eta)$. On the other hand, if $\theta^{(n)}(\eta) = \bar{\theta}$, it is trivial that $\theta^{(n)}(\eta') \leq \theta^{(n)}(\eta)$. Following a similar logic, when $T < T^\star$, we know that $\theta^{(n)}(\eta') \geq \theta^{(n)}(\eta)$. The desired result holds. □

Proof of Proposition 4

First, we know that under (31) and (32) there exists a unique Nash equilibrium. In the proof below we omit the contract parameter designations $(T, \boldsymbol{\xi})$ for notational simplicity.

a) Taking the cross derivative for $\pi_i^{(n)}(\boldsymbol{\theta})$ with respect to θ_i and θ_j, $i, j = 1, 2, j \neq i$, yields

$$\begin{aligned}
\frac{\partial^2\pi_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j} &= \frac{\partial^2\lambda_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\left[R_i(\theta_i) - C_i'(\mu_i)\right] + \frac{\partial\lambda_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_j}R_i'(\theta_i) - C_i''(\mu_i)\frac{\partial\lambda_i^{(n)}(\theta)}{\partial\theta_j}\left[\frac{\partial\lambda_i^{(n)}(\theta)}{\partial\theta_i} + 1\right] \\
&= \frac{\partial^2\lambda_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}[R_i(\theta_i) - C_i'(\mu_i)] - \frac{\partial\lambda_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_j}\left[C_i''(\mu_i)\left(\frac{\partial\lambda_i^{(n)}(\theta)}{\partial\theta_i} + 1\right) - R_i'(\theta_i)\right].
\end{aligned}$$

If C_i is linear (i.e. $C_i''(\mu_i) = 0$, $i = 1, 2$), then

$$\frac{\partial^2\pi_i^{(n)}}{\partial\theta_i\partial\theta_j} = \frac{\partial^2\lambda_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\left[R_i(\theta_i) - C_i'(\mu_i)\right] + \frac{\partial\lambda_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_j}R_i'(\theta_i) \leq 0,$$

where the inequality follows from Assumptions 2 and 3 and also from (31). In this case, $\pi_i^{(n)}$ is submodular in $\boldsymbol{\theta}$, which implies that the optimal-response service level of hospital i, $\theta_i(\theta_j)$ is decreasing in θ_j (Topkis 1998, Theorem 2.8.1). Hence, when hospital j exits the market, which is equivalent to reducing its service level to zero, the monotonicity of the optimal-response service level of hospital i implies that the $\theta_i^{(2)} \leq \theta_i^{(1)}$.

b) Suppose now that (48) is satisfied, so that $d_i > 0$ for $i = 1, 2$ and the cost functions for both hospitals are strictly convex. Then,

$$\begin{aligned}\frac{\partial^2 \pi_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j} &\geq -\delta_{ij}\left[\bar{R}_i - c_i\right] - \frac{\partial \lambda_i^{(n)}(\theta)}{\partial \theta_j}\left[C_i^{''}(\mu_i)\left(\frac{\partial \lambda_i^{(n)}(\theta)}{\partial \theta_i} + 1\right) - R_i'(\theta_i)\right] \\ &\geq -\delta_{ij}\left[\bar{R}_i - c_i\right] - \frac{\partial \lambda_i^{(n)}(\theta)}{\partial \theta_j}\left[d_i(\gamma_{ii}+1) - \bar{R}_i'\right] \\ &\geq -\delta_{ij}\left[\bar{R}_i - c_i\right] + \bar{\gamma}_{ij}\left[d_i(\gamma_{ii}+1) - \bar{R}_i'\right] \geq 0, \end{aligned} \tag{A26}$$

where the first inequality follows from (13) and (16), the second, from Assumption 2, (18), and (8), the third, from (11), and the last, from (48) and (49). Thus, $\pi_i^{(n)}(\boldsymbol{\theta})$ is supermodular in $\boldsymbol{\theta}$, which implies that the optimal-response service level of hospital i is increasing in θ_j (Topkis 1998, Theorem 2.8.1). Hence, when hospital j exits the market, which is equivalent to reducing its service level to zero, the monotonicity of the optimal-response service level of hospital i implies that the $\theta_i^{(2)} \geq \theta_i^{(1)}$. □

Proof of Proposition 5

a) The condition $\eta < \eta^\star(T)$ ensures the existence and uniqueness of a symmetric Nash equilibrium. Suppose that the cost function for any hospital is linear, i.e. $C(\mu_i) = c\mu_i$, where $\mu_i = \lambda_i^{(n)}(\boldsymbol{\theta}) + \theta_i$ is the service rate of i-th hospital. For convenience, define

$$\Psi_i^{(n)}(t) = \frac{\partial \pi_i(t,...,t)}{\partial \theta_i} = \frac{\partial \lambda_i^{(n)}(t,...,t)}{\partial \theta_i}[r - c - \eta r e^{-tT}] - c + \lambda_i^{(n)}(t,...,t) T\eta r e^{-tT}, \tag{A27}$$

where we have dropped the subscript i for the revenue and cost parameters. Taking the derivative of (A27) with respect to t, we have

$$\begin{aligned}\frac{d\Psi_i^{(n)}(t)}{dt} &= \left[\frac{\partial^2 \lambda_i^{(n)}(t,...,t)}{\partial \theta_i^2} + \sum_{j\neq i}\frac{\partial^2 \lambda_i^{(n)}(t,...,t)}{\partial\theta_i \partial\theta_j}\right][r - \eta r e^{-tT} - c] \\ &\quad - \left[\lambda_i^{(n)}(t,...,t)T - 2\frac{\partial \lambda_i^{(n)}(t,...,t)}{\partial \theta_i} - \sum_{j\neq i}\frac{\partial \lambda_i^{(n)}(t,...,t)}{\partial \theta_j}\right] T\eta r e^{-tT} \qquad \text{(A28)} \\ &< \left[\frac{\partial^2 \lambda_i^{(n)}(t,...,t)}{\partial \theta_i^2} + \sum_{j\neq i}\frac{\partial^2 \lambda_i^{(n)}(t,...,t)}{\partial\theta_i \partial\theta_j}\right][r - \eta r e^{-tT} - c] \\ &\quad + \left[2\frac{\partial \lambda_i^{(n)}(t,...,t)}{\partial \theta_i} + \sum_{j\neq i}\frac{\partial \lambda_i^{(n)}(t,...,t)}{\partial \theta_j}\right] T\eta r \qquad \text{(A29)} \\ &\leq -(\delta + \bar{\Delta})[r - \eta r e^{-tT} - c] + (2\bar{\gamma} - (n-1)\bar{\Gamma})T\eta r \qquad \text{(A30)} \\ &\leq -(\delta + \bar{\Delta})[r - \eta r - c] + (2\bar{\gamma} - (n-1)\bar{\Gamma})T\eta r \qquad \text{(A31)} \\ &< 0, \qquad \text{(A32)}\end{aligned}$$

where the first inequality is due to the assumption that $\frac{\partial \lambda_i^{(n)}(t,...,t)}{\partial \theta_i} \geq 0$ and $\frac{\partial \lambda_i^{(n)}(t,...,t)}{\partial \theta_i} + \sum_{j\neq i}\frac{\partial \lambda_i^{(n)}(t,...,t)}{\partial \theta_j} > 0$, the second, to the assumption that $\frac{\partial^2 \lambda_i^{(n)}(t,...,t)}{\partial \theta_i^2} \leq -\delta$, $\frac{\partial^2 \lambda_i^{(n)}(t,...,t)}{\partial \theta_i \theta_j} \leq -\bar{\Delta}$,

$\frac{\partial\lambda_i^{(n)}(t,...,t)}{\partial\theta_i} \le \bar{\gamma}$, and $\frac{\partial\lambda_i^{(n)}(t,...,t)}{\partial\theta_j} \le -\bar{\Gamma}$, and the last, to the assumption that $\eta < \frac{(\delta+\bar{\Delta})(r-c)}{(2\bar{\gamma}-(n-1)\bar{\Gamma})Tr+(\delta+\bar{\Delta})r}$. The strict monotonicity ensures that the symmetric equilibrium must be the smallest nonnegative value of θ such that $\Psi_i^{(n)}(\theta) \le 0$. If the Nash equilibrium service level is 0, the result of part a) obviously holds. Consider a strictly positive symmetric equilibrium, denoted by $\theta^{(n)}$ (here and in the following we drop the dependence on η and T). Then,

$$0 = \Psi_i^{(n)}\left(\theta^{(n)}\right) = \frac{\partial\lambda_i^{(n)}(\theta^{(n)},...,\theta^{(n)})}{\partial\theta_i}\left[r-c-\eta re^{-\theta^{(n)}T}\right] - c + \lambda_i^{(n)}(\theta^{(n)},...,\theta^{(n)})T\eta re^{-\theta^{(n)}T}. \quad \text{(A33)}$$

Without loss of generality, suppose that hospital n exits the market, i.e. $\theta_n = 0$. Setting $\theta_1 = ... = \theta_{n-1} = \theta^{(n)}$, for any $i < n$ we have

$$\begin{aligned}
\Psi_i^{(n-1)}\left(\theta^{(n)}\right) &= = \frac{\partial\pi_i^{(n-1)}\left(\theta^{(n)},...,\theta^{(n)}\right)}{\partial\theta_i} = \frac{\partial\pi_i^{(n)}\left(\theta^{(n)},...,\theta^{(n)},0\right)}{\partial\theta_i} \\
&= \frac{\partial\lambda_i^{(n)}\left(\theta^{(n)},...,\theta^{(n)},0\right)}{\partial\theta_i}\left[r-c-\eta re^{-\theta^{(n)}T}\right] - c + \lambda_i^{(n)}\left(\theta^{(n)},...,\theta^{(n)},0\right)T\eta re^{-\theta^{(n)}T} \\
&\ge \frac{\partial\lambda_i^{(n)}\left(\theta^{(n)},...,\theta^{(n)},\theta^{(n)}\right)}{\partial\theta_i}\left[r-c-\eta re^{-\theta^{(n)}T}\right] - c + \lambda_i^{(n)}\left(\theta^{(n)},...,\theta^{(n)},\theta^{(n)}\right)T\eta re^{-\theta^{(n)}T} \\
&= \Psi_i^{(n)}\left(\theta^{(n)}\right) = 0, \qquad \text{(A34)}
\end{aligned}$$

where the inequality is due to the submodularity of $\lambda_i^{(n)}(\boldsymbol{\theta})$ in (θ_i, θ_n) and the fact that $\lambda_i^{(n)}(\boldsymbol{\theta})$ is decreasing in θ_n.

Since $\Psi_i^{(n-1)}(t)$ is strictly increasing in t, (A34) implies that the symmetric Nash equilibrium service level for the oligopoly setting with $n-1$ hospitals, $\theta^{(n-1,n)}$ (i.e. the value that satisfies $\Psi_i^{(n-1)}\left(\theta^{(n-1,n)}\right) = 0$) cannot be less than $\theta^{(n)}$.

b) The proof is similar to that in part b) of Proposition 4. Specifically, under (54) the cost functions for all hospitals are strictly convex. Then,

$$\begin{aligned}
\frac{\partial^2\pi_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j} &= \frac{\partial^2\lambda_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}[r_i - \eta r_i e^{-\theta_i T} - C_i'(\mu_i)] - \frac{\partial\lambda_i^{(n)}(\boldsymbol{\theta})}{\partial\theta_j}\left[C_i''(\mu_i)\left(\frac{\partial\lambda_i^{(n)}(\theta)}{\partial\theta_i}+1\right) - \eta rTe^{-\theta_i T}\right] \\
&\ge -\delta_{ij}\left[r_i - c_i\right] - \frac{\partial\lambda_i^{(n)}(\theta)}{\partial\theta_j}\left[C_i''(\mu_i)\left(\frac{\partial\lambda_i^{(n)}(\theta)}{\partial\theta_i}+1\right) - \eta r_i Te^{-\theta_i T}\right] \\
&\ge -\delta_{ij}\left[r_i - c_i\right] - \frac{\partial\lambda_i^{(n)}(\theta)}{\partial\theta_j}\left[d_i(\gamma_{ii}+1) - \eta r_i T\right] \\
&\ge -\delta_{ij}\left[r_i - c_i\right] + \bar{\gamma}_{ij}\left[d_i(\gamma_{ii}+1) - \eta r_i T\right] \ge 0, \qquad \text{(A35)}
\end{aligned}$$

where the first inequality follows from (13) and (16), the second, from Assumption 2, (18), and (8), the third, from (11), and the last, from (54) and (55). Thus, $\pi_i^{(n)}(\boldsymbol{\theta})$ is supermodular in $\boldsymbol{\theta}$. As Theorem 2.8.1 in Topkis (1998) implies, the optimal-response service level of hospital i is increasing in the service level of hospital j, θ_j. Thus, when hospital j exits the market, the monotonicity of the optimal-response service level of hospital i implies that the $\theta_i^{(n)}(\eta,T) \ge \theta_i^{(n-1,j)}(\eta,T)$. □

Proof of Lemma 1

When the service level of the competitor of hospital $i=1,2$ is set at $\theta_j, j\neq i$, the profit function for hospital i is given by

$$\begin{aligned}\pi_i^{(2)}(\boldsymbol{\theta}) &= r_i\left(1+\eta\left(\theta_i-\frac{1}{T}\right)\right)\left(\alpha_i\theta_i-\rho_{ij}\theta_i\theta_j\right)-b_i\left((\alpha_i+1)\theta_i-\rho_{ij}\theta_i\theta_j\right)^2\\ &= r_i\left(1+\eta\left(\theta_i-\frac{1}{T}\right)\right)\left(\alpha_i-\rho_{ij}\theta_j\right)\theta_i-b_i\left((\alpha_i+1)-\rho_{ij}\theta_j\right)^2\theta_i^2\\ &= r_i\left(1-\frac{\eta}{T}\right)\left(\alpha_i-\rho_{ij}\theta_j\right)\theta_i+\left(r_i\eta\left(\alpha_i-\rho_{ij}\theta_j\right)-b_i\left((\alpha_i+1)-\rho_{ij}\theta_j\right)^2\right)\theta_i^2. \qquad \text{(A36)}\end{aligned}$$

Thus,

$$\frac{\partial\pi_i^{(2)}}{\partial\theta_i}=r_i\left(1-\frac{\eta}{T}\right)\left(\alpha_i-\rho_{ij}\theta_j\right)+2\left(\eta r_i\left(\alpha_i-\rho_{ij}\theta_j\right)-b_i\left((\alpha_i+1)-\rho_{ij}\theta_j\right)^2\right)\theta_i. \qquad \text{(A37)}$$

Note that $\theta_i^{\mathrm{s}}(\theta_j)$ is the solution to $\frac{\partial\pi_i^{(2)}}{\partial\theta_i}=0$. As (A37) implies, $\pi_i^{(2)}(\boldsymbol{\theta})$ is a quasiconcave function of θ_i under (63) and (62). Indeed, $\frac{\partial\pi_i^{(2)}}{\partial\theta_i}$ is positive at $\theta_i=0$. If $\eta r_i\left(\alpha_i-\rho_{ij}\theta_j\right)-b_i\left((\alpha_i+1)-\rho_{ij}\theta_j\right)^2\geq 0$, $\pi_i^{(2)}$ is monotone increasing in θ_i and is quasiconcave. If, on the other hand, $\eta r_i\left(\alpha_i-\rho_{ij}\theta_j\right)-b_i\left((\alpha_i+1)-\rho_{ij}\theta_j\right)^2<0$, $\pi_i^{(2)}$ is strictly concave in θ_i. Thus, the optimal response of hospital i is either $\bar{\theta}_i$ or $\theta_i^{\mathrm{s}}(\theta_j)$, whichever value is lower. □

Proof of Proposition 6

Note that if (67) holds, then $\eta<T$. Since $\pi_i^{(2)}(\boldsymbol{\theta})$ is quasiconcave in θ_i and $\pi_j^{(2)}(\boldsymbol{\theta})$ is quasiconcave in θ_j, there exists a Nash equilibrium, as follows from Theorem 1 in Cachon and Netessine (2004).

The uniqueness of the Nash equilibrium is guaranteed if the profit functions of the competing hospitals satisfy a strict "diagonal dominance" condition:

$$\left|\frac{\partial^2\pi_i(\theta_i,\theta_j,T,\eta)}{\partial\theta_i^2}\right|>\left|\frac{\partial^2\pi_i(\theta_i,\theta_j,T,\eta)}{\partial\theta_i\partial\theta_j}\right|,\theta_i\in\left[0,\bar{\theta}_i\right],i,j=1,2,j\neq i. \qquad \text{(A38)}$$

Note that

$$\begin{aligned}\frac{\partial^2\pi_i(\theta_i,\theta_j,T,\eta)}{\partial\theta_i^2} &= -2b_i\left(\left(\alpha_i-\rho_{ij}\theta_j+1-\frac{\eta r_i}{2b_i}\right)^2+\left(1-\frac{\eta r_i}{4b_i}\right)\frac{\eta r_i}{b_i}\right)\\ &\leq -2b_i\left(\left(\alpha_i-\rho_{ij}\theta_j+1-\frac{\eta r_i}{2b_i}\right)^2-\left(\frac{\eta r_i}{2b_i}\right)^2\right)\\ &= -2b_i\left(\alpha_i-\rho_{ij}\theta_j+1\right)\left(\alpha_i-\rho_{ij}\theta_j+1-\frac{\eta r_i}{b_i}\right)<0,\ i,j=1,2,j\neq i, \qquad \text{(A39)}\end{aligned}$$

where the last inequality follows from $\eta<\frac{b_i}{r_i}(1+\alpha_i)$ and from $\bar{\theta}_j<\frac{1}{\rho_{ij}}\left(\alpha_i+1-\frac{\eta r_i}{b_i}\right)$ (both implied by (69)).

The partial cross-derivative of the profit function is given by

$$\begin{aligned}&\frac{\partial^2\pi_i\left(\theta_i,\theta_j,T,\eta\right)}{\partial\theta_i\partial\theta_j}\\&=\rho_{ij}\left(-r_i\left(1-\frac{\eta}{T}+2\eta\theta_i\right)+4b_i\theta_i\left(1+\alpha_i-\rho_{ij}\theta_j\right)\right),\end{aligned}\tag{A40}$$

so that we have

$$\begin{aligned}&\left|\frac{\partial^2\pi_i\left(\theta_i,\theta_j,T,\eta\right)}{\partial\theta_i^2}\right|-\left|\frac{\partial^2\pi_i\left(\theta_i,\theta_j,T,\eta\right)}{\partial\theta_i\partial\theta_j}\right|\\&\geq 2b_i\left(\left(\alpha_i-\rho_{ij}\theta_j+1-\frac{\eta r_i}{2b_i}\right)^2-\left(\frac{\eta r_i}{2b_i}\right)^2\right)\\&-\rho_{ij}\left|-r_i\left(1-\frac{\eta}{T}+2\eta\theta_i\right)+4b_i\theta_i\left(1+\alpha_i-\rho_{ij}\theta_j\right)\right|.\end{aligned}\tag{A41}$$

Note that for $\theta_i\in\left[0,\bar{\theta}_i\right]$,

$$\begin{aligned}&\left|-r_i\left(1-\frac{\eta}{T}+2\eta\theta_i\right)+4b_i\theta_i\left(1+\alpha_i-\rho_{ij}\theta_j\right)\right|\\&=\left|-r_i\left(1-\frac{\eta}{T}\right)+4b_i\theta_i\left(1+\alpha_i-\rho_{ij}\theta_j-\frac{\eta r_i}{2b_i}\right)\right|\\&=\max\left(r_i\left(1-\frac{\eta}{T}\right),\left|-r_i\left(1-\frac{\eta}{T}\right)+4b_i\bar{\theta}_i\left(1+\alpha_i-\rho_{ij}\theta_j-\frac{\eta r_i}{2b_i}\right)\right|\right),\end{aligned}\tag{A42}$$

since $-r_i\left(1-\frac{\eta}{T}\right)+4b_i\theta_i\left(1+\alpha_i-\rho_{ij}\theta_j-\frac{\eta r_i}{2b_i}\right)$ is a linear function of θ_i.

The right-hand side of (A41) is positive if

$$\begin{aligned}&2b_i\left(\left(\alpha_i-\rho_{ij}\theta_j+1-\frac{\eta r_i}{2b_i}\right)^2-\left(\frac{\eta r_i}{2b_i}\right)^2\right)\\&-\rho_{ij}\max\left(r_i\left(1-\frac{\eta}{T}\right),\left|-r_i\left(1-\frac{\eta}{T}\right)+4b_i\bar{\theta}_i\left(1+\alpha_i-\rho_{ij}\theta_j-\frac{\eta r_i}{2b_i}\right)\right|\right)>0.\end{aligned}\tag{A43}$$

For (A43) to hold we need, specifically, that

$$2b_i\left(\left(\alpha_i-\rho_{ij}\theta_j+1-\frac{\eta r_i}{2b_i}\right)^2-\left(\frac{\eta r_i}{2b_i}\right)^2\right)-\rho_{ij}r_i\left(1-\frac{\eta}{T}\right)>0,\tag{A44}$$

which is ensured if

$$2b_i\left(\left(\alpha_i-\rho_{ij}\bar{\theta}_j+1-\frac{\eta r_i}{2b_i}\right)^2-\left(\frac{\eta r_i}{2b_i}\right)^2\right)-\rho_{ij}r_i\left(1-\frac{\eta}{T}\right)>0.\tag{A45}$$

(A45) follows from

$$\eta<\frac{b}{r}\left(1+\alpha\right)\tag{A46}$$

and

$$\bar{\theta}_j<\frac{1}{\rho_{ij}}\left(1+\alpha_i-\frac{\eta r_i}{2b_i}-\sqrt{\left(\frac{\eta r_i}{2b_i}\right)^2+\left(\frac{\rho_{ij}r_i}{2b_i}\right)\left(1-\frac{\eta}{T}\right)}\right),\tag{A47}$$

or (68).

In order to ensure that (A43) holds we need that

$$2b_i\left(\left(\alpha_i-\rho_{ij}\theta_j+1-\frac{\eta r_i}{2b_i}\right)^2-\left(\frac{\eta r_i}{2b_i}\right)^2\right)>\rho_{ij}\left(-r_i\left(1-\frac{\eta}{T}\right)+4b_i\bar{\theta}_i\left(1+\alpha_i-\rho_{ij}\theta_j-\frac{\eta r_i}{2b_i}\right)\right) \quad \text{(A48)}$$

and

$$2b_i\left(\left(\alpha_i-\rho_{ij}\theta_j+1-\frac{\eta r_i}{2b_i}\right)^2-\left(\frac{\eta r_i}{2b_i}\right)^2\right)>\rho_{ij}\left(r_i\left(1-\frac{\eta}{T}\right)-4b_i\bar{\theta}_i\left(1+\alpha_i-\rho_{ij}\theta_j-\frac{\eta r_i}{2b_i}\right)\right). \quad \text{(A49)}$$

Note that (A49) follows from (A44). Focusing on (A48), we can express it as

$$z_i^2-2\rho_{ij}\bar{\theta}_i z_i+\frac{r_i\rho_{ij}}{2b_i}\left(1-\frac{\eta}{T}\right)-\left(\frac{\eta r_i}{2b_i}\right)^2>0, \quad \text{(A50)}$$

where $z_i=\alpha_i-\rho_{ij}\theta_j+1-\frac{\eta r_i}{2b_i}$, so that $\alpha_i-\rho_{ij}\bar{\theta}_j+1-\frac{\eta r_i}{2b_i}\le z_i\le\alpha_i+1-\frac{\eta r_i}{2b_i}$. Evaluating the derivative of the left-hand side of (A50) with respect to z_i at the lower-bound value $\alpha_i-\rho_{ij}\bar{\theta}_j+1-\frac{\eta r_i}{2b_i}$, we get

$$2\left(\alpha_i-\rho_{ij}\bar{\theta}_j-\rho_{ij}\bar{\theta}_i+1-\frac{\eta r_i}{2b_i}\right), \quad \text{(A51)}$$

which is positive under (69). Since the derivative of the left-hand side of (A50) at $z_i=\alpha_i-\rho_{ij}\bar{\theta}_j+1-\frac{\eta r_i}{2b_i}$ is positive, the left-hand side of (A50) is positive on $\alpha_i-\rho_{ij}\bar{\theta}_j+1-\frac{\eta r_i}{2b_i}\le z_i\le\alpha_i+1-\frac{\eta r_i}{2b_i}$ if $\alpha_i-\rho_{ij}\bar{\theta}_j+1-\frac{\eta r_i}{2b_i}$ is greater than the larger root of $z_i^2-2\rho_{ij}\bar{\theta}_i z_i+\frac{r_i\rho_{ij}}{2b_i}\left(1-\frac{\eta}{T}\right)-\left(\frac{\eta r_i}{2b_i}\right)^2$, if it exists. This largest root, if it exists, is given by

$$\rho_{ij}\bar{\theta}_i+\sqrt{\left(\rho_{ij}\bar{\theta}_i\right)^2+\left(\frac{\eta r_i}{2b_i}\right)^2-\frac{r_i\rho_{ij}}{2b_i}\left(1-\frac{\eta}{T}\right)}. \quad \text{(A52)}$$

Note that (69) implies

$$\alpha_i-\rho_{ij}\bar{\theta}_j-\rho_{ij}\bar{\theta}_i+1-\frac{\eta r_i}{2b_i}>\frac{\eta r_i}{2b_i}+\rho_{ij}\bar{\theta}_i>\sqrt{\left(\frac{\eta r_i}{2b_i}\right)^2+\left(\rho_{ij}\bar{\theta}_i\right)^2}. \quad \text{(A53)}$$

Thus, under (69),

$$\alpha_i-\rho_{ij}\bar{\theta}_j+1-\frac{\eta r_i}{2b_i}>\rho_{ij}\bar{\theta}_i+\sqrt{\left(\frac{\eta r_i}{2b_i}\right)^2+\left(\rho_{ij}\bar{\theta}_i\right)^2}>\rho_{ij}\bar{\theta}_i+\sqrt{\left(\rho_{ij}\bar{\theta}_i\right)^2+\left(\frac{\eta r_i}{2b_i}\right)^2-\frac{r_i\rho_{ij}}{2b_i}\left(1-\frac{\eta}{T}\right)}, \quad \text{(A54)}$$

and (A48) holds. □

Proof of Proposition 7

As Proposition 6 states, under (62) and (66)-(68) there exists a unique Nash equilibrium. From the symmetric version of (64) we have the following expression for a symmetric Nash equilibrium service level in the absence of any constraints on the service level values:

$$\hat{\theta}=\frac{(\alpha-\rho\hat{\theta})(1-\eta/T)\left(\frac{r}{b}\right)}{2\left[\left(1+\alpha-\rho\hat{\theta}\right)^2-(\alpha-\rho\hat{\theta})\eta\left(\frac{r}{b}\right)\right]}. \quad \text{(A55)}$$

Denoting $\hat{x} = \alpha - \rho\hat{\theta}$, we can express (A55) as

$$\hat{x} = \alpha - \frac{\hat{x}\left(\frac{\rho r}{2b}\right)\left(1-\frac{\eta}{T}\right)}{\left(\left(\hat{x}+1\right)^2 - \hat{x}\left(\frac{\eta r}{b}\right)\right)}, \tag{A56}$$

or

$$\left(\hat{x}\right)^3 + \left(2-\alpha-\frac{\eta r}{b}\right)\left(\hat{x}\right)^2 + \left(1-2\alpha+\frac{\rho r}{2b}+\frac{\eta r}{b}\left(\alpha-\frac{\rho}{2T}\right)\right)\hat{x} - \alpha = 0. \tag{A57}$$

We can express (A57) as

$$\left(\hat{x}\right)^3 + p\left(\hat{x}\right)^2 + q\hat{x} - \alpha = 0, \tag{A58}$$

where

$$p = 2-\alpha-\frac{\eta r}{b}, \tag{A59}$$

$$q = 1-2\alpha+\frac{\rho r}{2b}+\frac{\eta r}{b}\left(\alpha-\frac{\rho}{2T}\right). \tag{A60}$$

The left-hand side of (A58) can be expressed as

$$\begin{aligned}
&\left(\hat{x}+\frac{p}{3}\right)^3 - 3\hat{x}\left(\frac{p}{3}\right)^2 - \left(\frac{p}{3}\right)^3 + q\hat{x} - \alpha \\
&= \left(\hat{x}+\frac{p}{3}\right)^3 + \hat{x}\left(q-\frac{p^2}{3}\right) - \alpha - \left(\frac{p}{3}\right)^3 = \left(\hat{x}+\frac{p}{3}\right)^3 + \left(\hat{x}+\frac{p}{3}\right)\left(q-\frac{p^2}{3}\right) - \frac{p}{3}\left(q-\frac{p^2}{3}\right) - \alpha - \left(\frac{p}{3}\right)^3 \\
&= \left(\hat{x}+\frac{p}{3}\right)^3 + \left(\hat{x}+\frac{p}{3}\right)\left(q-\frac{p^2}{3}\right) - \alpha - \frac{pq}{3} + \frac{2p^3}{27} \\
&= t^3 + ut + v,
\end{aligned} \tag{A61}$$

with $t = \hat{x} + \frac{p}{3}$ and

$$\begin{aligned}
u = q - \frac{p^2}{3} &= 1-2\alpha+\frac{\rho r}{2b}+\frac{\eta r}{b}\left(\alpha-\frac{\rho}{2T}\right) - \frac{1}{3}\left(2-\alpha-\frac{\eta r}{b}\right)^2 \\
&= 1-2\alpha+\frac{\rho r}{2b} - \frac{1}{3}\left(2-\alpha\right)^2 + \frac{\eta r}{b}\left(\alpha-\frac{\rho}{2T}+\frac{2}{3}\left(2-\alpha\right)\right) - \frac{1}{3}\left(\frac{\eta r}{b}\right)^2 \\
&= \frac{\rho r}{2b} - \frac{\left(\alpha+1\right)^2}{3} + \frac{\eta r}{b}\left(\alpha-\frac{\rho}{2T}+\frac{2}{3}\left(2-\alpha\right)\right) - \frac{1}{3}\left(\frac{\eta r}{b}\right)^2 \\
&= \frac{\rho r}{2b} - 3\left(\frac{\alpha+1}{3}\right)^2 + \frac{\eta r}{b}\left(\frac{\alpha+1}{3}+1-\frac{\rho}{2T}\right) - \frac{1}{3}\left(\frac{\eta r}{b}\right)^2
\end{aligned} \tag{A62}$$

and

$$\begin{aligned}
v &= -\alpha - \frac{pq}{3} + \frac{2p^3}{27} = -\alpha - \frac{1}{3}\left(2-\alpha-\frac{\eta r}{b}\right)\left(1-2\alpha+\frac{\rho r}{2b}+\frac{\eta r}{b}\left(\alpha-\frac{\rho}{2T}\right)\right) + \frac{2}{27}\left(2-\alpha-\frac{\eta r}{b}\right)^3 \\
&= -\alpha - \frac{1}{3}\left(2-\alpha\right)\left(1-2\alpha+\frac{\rho r}{2b}\right) + \frac{2}{27}\left(2-\alpha\right)^3 \\
&+ \left(\frac{\eta r}{b}\right)\left(\frac{1}{3}\left(1-2\alpha+\frac{\rho r}{2b}\right) - \frac{1}{3}\left(2-\alpha\right)\left(\alpha-\frac{\rho}{2T}\right) - \frac{2}{9}\left(2-\alpha\right)^2\right) \\
&+ \left(\frac{\eta r}{b}\right)^2\left(\frac{1}{3}\left(\alpha-\frac{\rho}{2T}\right) + \frac{2}{9}\left(2-\alpha\right)\right) - \frac{2}{27}\left(\frac{\eta r}{b}\right)^3.
\end{aligned} \tag{A63}$$

Note that

$$\begin{aligned}
&-\alpha-\frac{1}{3}(2-\alpha)\left(1-2\alpha+\frac{\rho r}{2b}\right)+\frac{2}{27}(2-\alpha)^3\\
&=-(\alpha+1)+1-\frac{1}{3}(3-(1+\alpha))\left(3-2(1+\alpha)+\frac{\rho r}{2b}\right)+\frac{2}{27}(3-(1+\alpha))^3\\
&=1-3-\frac{\rho r}{2b}+2+(\alpha+1)\left(-1+\frac{1}{3}\left(3+\frac{\rho r}{2b}\right)+2-2\right)+(\alpha+1)^2\left(-\frac{2}{3}+\frac{2}{3}\right)-\frac{2}{27}(1+\alpha)^3\\
&=\frac{\rho r}{2b}\left(\frac{\alpha+1}{3}-1\right)-2\left(\frac{\alpha+1}{3}\right)^3
\end{aligned}\tag{A64}$$

and

$$\begin{aligned}
&\frac{1}{3}\left(1-2\alpha+\frac{\rho r}{2b}\right)-\frac{1}{3}(2-\alpha)\left(\alpha-\frac{\rho}{2T}\right)-\frac{2}{9}(2-\alpha)^2\\
&=\frac{1}{3}\left(3-2(1+\alpha)+\frac{\rho r}{2b}\right)-\frac{1}{3}(3-(1+\alpha))\left((\alpha+1)-1-\frac{\rho}{2T}\right)-\frac{2}{9}(3-(1+\alpha))^2\\
&=1+\frac{\rho r}{6b}+1+\frac{\rho}{2T}-2+(\alpha+1)\left(-\frac{2}{3}-1-\frac{1}{3}\left(1+\frac{\rho}{2T}\right)+\frac{4}{3}\right)+\frac{1}{9}(\alpha+1)^2\\
&=\frac{\rho r}{6b}+\frac{\rho}{2T}+(\alpha+1)\left(-\frac{2}{3}-\frac{\rho}{6T}\right)-\frac{2}{9}(\alpha+1)^2\\
&=\frac{\rho r}{6b}+\frac{\rho}{2T}-\left(\frac{\alpha+1}{3}\right)\left(2+\frac{\rho}{2T}\right)+\left(\frac{\alpha+1}{3}\right)^2.
\end{aligned}\tag{A65}$$

Finally,

$$\begin{aligned}
&\frac{1}{3}\left(\alpha-\frac{\rho}{2T}\right)+\frac{2}{9}(2-\alpha)=\frac{1}{3}\left(\alpha+1-1-\frac{\rho}{2T}\right)+\frac{2}{9}(3-(1+\alpha))\\
&=\frac{1}{3}\left(1-\frac{\rho}{2T}\right)+\frac{1}{9}(\alpha+1),
\end{aligned}\tag{A66}$$

so that (A63) becomes

$$\begin{aligned}
v=&\frac{\rho r}{2b}\left(\frac{\alpha+1}{3}-1\right)-2\left(\frac{\alpha+1}{3}\right)^3\\
&+\left(\frac{\eta r}{b}\right)\left(\frac{\rho r}{6b}+\frac{\rho}{2T}-\left(\frac{\alpha+1}{3}\right)\left(2+\frac{\rho}{2T}\right)+\left(\frac{\alpha+1}{3}\right)^2\right)\\
&+\left(\frac{\eta r}{b}\right)^2\left(\frac{1}{3}\left(1-\frac{\rho}{2T}\right)+\frac{1}{9}(\alpha+1)\right)-\frac{2}{27}\left(\frac{\eta r}{b}\right)^3.
\end{aligned}\tag{A67}$$

The only real root of the equation

$$t^3+ut+v=0\tag{A68}$$

is

$$\hat{t}=\left(\sqrt{\left(\frac{u}{3}\right)^3+\left(\frac{v}{2}\right)^2}-\frac{v}{2}\right)^{\frac{1}{3}}-\left(\sqrt{\left(\frac{u}{3}\right)^3+\left(\frac{v}{2}\right)^2}+\frac{v}{2}\right)^{\frac{1}{3}}.\tag{A69}$$

Defining

$$\hat{u} = \frac{u}{3} = \frac{\rho r}{6b} - \left(\frac{\alpha+1}{3}\right)^2 + \left(\frac{\eta r}{3b}\right)\left(\frac{\alpha+1}{3} + 1 - \frac{\rho}{2T}\right) - \left(\frac{\eta r}{3b}\right)^2, \tag{A70}$$

$$\begin{aligned}\hat{v} = \frac{v}{2} = {} & \frac{\rho r}{4b}\left(\frac{\alpha+1}{3} - 1\right) - \left(\frac{\alpha+1}{3}\right)^3 \\ & + \frac{3}{2}\left(\frac{\eta r}{3b}\right)\left(\frac{\rho r}{6b} + \frac{\rho}{2T} - \left(\frac{\alpha+1}{3}\right)\left(2 + \frac{\rho}{2T}\right) + \left(\frac{\alpha+1}{3}\right)^2\right) \\ & + \frac{3}{2}\left(\frac{\eta r}{3b}\right)^2\left(\left(1 - \frac{\rho}{2T}\right) + \left(\frac{\alpha+1}{3}\right)\right) - \left(\frac{\eta r}{3b}\right)^3,\end{aligned} \tag{A71}$$

we get

$$\hat{t} = \left(\sqrt{(\hat{u})^3 + (\hat{v})^2} - \hat{v}\right)^{\frac{1}{3}} - \left(\sqrt{(\hat{u})^3 + (\hat{v})^2} + \hat{v}\right)^{\frac{1}{3}}. \tag{A72}$$

Note that the solution to (A55) is connected to $\hat{t}$ as follows:

$$\hat{\theta} = \frac{1}{\rho}\left(\alpha - \hat{t} + \frac{p}{3}\right). \tag{A73}$$

Since

$$\alpha + \frac{p}{3} = \alpha + \frac{1}{3}\left(2 - \alpha - \frac{\eta r}{b}\right) = \frac{2}{3}(\alpha+1) - \frac{1}{3}\left(\frac{\eta r}{b}\right), \tag{A74}$$

we get (90). Next, we prove that when $\hat{\theta} \geq \bar{\theta}$, $(\bar{\theta}, \bar{\theta})$ is the unique Nash equilibrium.

As follows from (A37), the symmetric equilibrium (θ, θ) satisfies the following equation

$$\frac{\partial \pi_i^{(2)}}{\partial \theta_i}(\theta, \theta) = r\left(1 - \frac{\eta}{T}\right)(\alpha - \rho\theta) + 2\left(\eta r\left(\alpha - \rho\theta\right) - b\left((\alpha+1) - \rho\theta\right)^2\right)\theta = 0, i = 1, 2. \tag{A75}$$

Note that $\lim_{\theta\to\infty} \frac{\partial \pi_i^{(2)}}{\partial \theta_i}(\theta, \theta) = -\infty$. Recall that for a given η, (A75) has a unique solution, $\hat{\theta}$. Thus, $\frac{\partial \pi_i^{(2)}}{\partial \theta_i}(\theta, \theta)$ "crosses" the zero value only once. Moreover, when $\theta < \hat{\theta}$, $\frac{\partial \pi_i^{(2)}}{\partial \theta_i}(\theta, \theta) > 0$ and when $\theta > \hat{\theta}$, $\frac{\partial \pi_i^{(2)}}{\partial \theta_i}(\theta, \theta) < 0$. For convenience, let us denote these results, collectively, as the "single-crossing property."

Consider the case of $\hat{\theta} \geq \bar{\theta}$ and suppose that there exist θ_1 and θ_2 such that (θ_1, θ_2) is a Nash equilibrium. Then, it is not possible that $\theta_1 = \theta_2$ unless $\theta_1 = \theta_2 = \hat{\theta}$ because (A75) has a unique solution $\hat{\theta}$. Without loss of generality, we assume that $0 < \theta_1 < \theta_2 \leq \bar{\theta} < \hat{\theta}$. Since (θ_1, θ_2) is a Nash equilibrium, we must have

$$\frac{\partial \pi_i^{(2)}}{\partial \theta_i}(\theta_1, \theta_2) = 0, i = 1, 2, \tag{A76}$$

$$\eta r\left(\alpha - \rho\theta_2\right) - b\left((\alpha+1) - \rho\theta_2\right)^2 < 0. \tag{A77}$$

Then, using (A76)–(A77) and the single-crossing property, we get

$$0 = \frac{\partial \pi_1^{(2)}}{\partial \theta_1}(\theta_1, \theta_2)$$

$$
\begin{aligned}
&= r\left(1-\frac{\eta}{T}\right)(\alpha-\rho\theta_2)+2\left(\eta r\left(\alpha-\rho\theta_2\right)-b\left((\alpha+1)-\rho\theta_2\right)^2\right)\theta_1\\
&> r\left(1-\frac{\eta}{T}\right)(\alpha-\rho\theta_2)+2\left(\eta r\left(\alpha-\rho\theta_2\right)-b\left((\alpha+1)-\rho\theta_2\right)^2\right)\theta_2\\
&= \frac{\partial\pi_2^{(2)}}{\partial\theta_2}(\theta_2,\theta_2) > \frac{\partial\pi_2^{(2)}}{\partial\theta_2}(\hat{\theta},\hat{\theta})=0,
\end{aligned}
\tag{A78}
$$

which is a contradiction. Thus, the only candidate for the Nash equilibrium is $(\bar{\theta},\bar{\theta})$. Indeed, $\bar{\theta}$ is the best response for hospital i when the service level for hospital j is $\bar{\theta}$ since

$$
\frac{\partial\pi_i^{(2)}}{\partial\theta_i}(\bar{\theta},\bar{\theta}) \geq \frac{\partial\pi_i^{(2)}}{\partial\theta_i}(\hat{\theta},\hat{\theta})=0. \tag{A79}
$$

Therefore, $(\bar{\theta},\bar{\theta})$ is a unique Nash equilibrium. □

Proof of Proposition 8

(a) Note that the Nash equilibrium satisfies the following reformulated best-response equation:

$$
\begin{aligned}
\theta^{\mathrm{NE}}(\eta) &= \frac{\left(\alpha-\rho\theta^{\mathrm{NE}}(\eta)\right)\left(1-\frac{\eta}{T}\right)\frac{r}{b}}{2\left(-\left(\alpha-\rho\theta^{\mathrm{NE}}(\eta)\right)\frac{r\eta}{b}+\left(\alpha+1-\rho\theta^{\mathrm{NE}}(\eta)\right)^2\right)}\\
&= \frac{\left(1-\frac{\eta}{T}\right)\frac{r}{b}}{2\left(-\frac{r\eta}{b}+\left(2+\left(\alpha-\rho\theta^{\mathrm{NE}}(\eta)\right)+\frac{1}{\left(\alpha-\rho\theta^{\mathrm{NE}}(\eta)\right)}\right)\right)}.
\end{aligned}
\tag{A80}
$$

If $\alpha-\rho\bar{\theta}\geq 1$, then $\alpha\geq 1$ and $\alpha-\rho x+\frac{1}{\alpha-\rho x}$ is decreasing in x over $[0,\bar{\theta}]$. This shows that

$$
\alpha+\frac{1}{\alpha} > \alpha-\rho\theta^{\mathrm{NE}}(\eta)+\frac{1}{\alpha-\rho\theta^{\mathrm{NE}}(\eta)}. \tag{A81}
$$

It follows from (A80) and the fact that $\theta^{\mathrm{NE}}(\eta)\in[0,\bar{\theta}]$ that

$$
\begin{aligned}
\theta^{\mathrm{NE}}(\eta) &= \frac{\left(1-\frac{\eta}{T}\right)\frac{r}{b}}{2\left(-\frac{r\eta}{b}+\left(2+\left(\alpha-\rho\theta^{\mathrm{NE}}(\eta)\right)+\frac{1}{\left(\alpha-\rho\theta^{\mathrm{NE}}(\eta)\right)}\right)\right)}\\
&\geq \frac{\left(1-\frac{\eta}{T}\right)\frac{r}{b}}{2\left(-\frac{r\eta}{b}+\left(2+\alpha+\frac{1}{\alpha}\right)\right)}\\
&= \theta^{\mathrm{M}}(\eta).
\end{aligned}
\tag{A82}
$$

Therefore, we have shown that $\theta^{\mathrm{NE}}(\eta)\geq\theta^{\mathrm{M}}(\eta)$, i.e. competition increases service levels.

(b) If $\alpha\leq 1$, then $\alpha-\rho\theta^{\mathrm{M}}(\eta)\leq 1$ and $\alpha-\rho x+\frac{1}{\alpha-\rho x}$ is increasing in x over $[0,\bar{\theta}]$. This shows that

$$
\alpha+\frac{1}{\alpha} < \alpha-\rho\theta^{\mathrm{NE}}(\eta)+\frac{1}{\alpha-\rho\theta^{\mathrm{NE}}(\eta)}. \tag{A83}
$$

It follows from (A80) and the fact that $\theta^{\mathrm{NE}}(\eta)\in[0,\bar{\theta}]$ that

$$
\begin{aligned}
\theta^{\mathrm{NE}}(\eta) &= \frac{\left(1-\frac{\eta}{T}\right)\frac{r}{b}}{2\left(-\frac{r\eta}{b}+\left(2+\left(\alpha-\rho\theta^{\mathrm{NE}}(\eta)\right)+\frac{1}{\left(\alpha-\rho\theta^{\mathrm{NE}}(\eta)\right)}\right)\right)}\\
&\leq \frac{\left(1-\frac{\eta}{T}\right)\frac{r}{b}}{2\left(-\frac{r\eta}{b}+\left(2+\alpha+\frac{1}{\alpha}\right)\right)}\\
&= \theta^{\mathrm{M}}(\eta).
\end{aligned}
\tag{A84}
$$

Therefore, we have shown that $\theta^{\mathrm{NE}}(\eta) \leq \theta^{\mathrm{M}}(\eta)$, i.e. competition decreases service levels. □

Lemma A1

LEMMA A1. *Consider the payer problem in the symmetric duopoly setting. Assume that (62) and (66)–(69) hold. The payer problem is equivalent to the following optimization problem:*

$$\min_{T,\eta} \quad 2r\left(\alpha\theta^{\mathrm{NE}} - \rho\theta^{\mathrm{NE}}\theta^{\mathrm{NE}}\right)\left(1+\eta\left(\theta^{\mathrm{NE}} - \frac{1}{T}\right)\right) \tag{A85}$$

$$\text{s.t.} \quad \eta \geq \eta^{\min}(T) \equiv \frac{2b}{r}\frac{\left((\alpha+1)-\frac{\rho}{T}\right)^2}{\left(\alpha-\frac{\rho}{T}\right)} - T \tag{A86}$$

$$\eta \leq \frac{-r\left(\alpha-\rho\bar{\theta}\right)+2b\left((\alpha+1)-\rho\bar{\theta}\right)^2\bar{\theta}}{r\left(\alpha-\rho\bar{\theta}\right)\left(2\bar{\theta}-\frac{1}{T}\right)} \tag{A87}$$

$$T_l \leq T \leq T_h, \tag{A88}$$

$$0 \leq \eta \leq \eta^{\max}(T). \tag{A89}$$

Furthermore, constraints (A86) and (A87) are compatible.

The equivalence result in Lemma A1 is significant because in the reformulation simple and explicit constraints replace implicit and complicated service-level and participation constraints. We show that the service-level constraint is replaced by (A86) and the participation constraint is automatically satisfied. Moreover, the inclusion of the redundant constraint (A87) does not alter the optimal solution for the payer problem but does allow us to develop monotone properties of the objective function for the payer problem. It is trivial to see that the payer problem in the monopoly setting can be obtained from the payer problem defined in Lemma A1 by letting $\rho = 0$ and replacing θ^{NE} with θ^{M}. It is important that constraints (A86) and (A87) are compatible. Otherwise, the feasible region for the equivalent payer problem is empty. On the other hand, constraints (A86) and (A89) are not always compatible.

Proof of Lemma A1

First, we show that the participation constraint is automatically satisfied. For any given service level θ_j for hospital j, the profit function for hospital i is equal to zero at $\theta_i = 0$. Therefore, the profit function for hospital i is greater than or equal to zero at the best response. For any η, the best response may not be greater than the minimum service level $\frac{1}{T}$ that the payer intends to impose. However, we prove that the payer can force hospital i to meet the minimum service level is equivalent to choosing η that is greater than or equal to a threshold. Therefore, once this threshold for η is imposed, the participation constraint is automatically satisfied.

Second, we show that the service level constraint can be replaced by (A86). Since $\bar{\theta} \geq \frac{1}{T}$, constraint (100) is equivalent to $\hat{\theta} \geq \frac{1}{T}$. As follows from the single-crossing property, constraint (100) is equivalent to the following inequality:

$$\begin{aligned}
0 &= \frac{\partial \pi_i^{(2)}}{\partial \theta_i}(\hat{\theta}, \hat{\theta}) \\
&\leq \frac{\partial \pi_i^{(2)}}{\partial \theta_i}(\frac{1}{T}, \frac{1}{T}) \\
&= r\left(1-\frac{\eta}{T}\right)\left(\alpha-\rho\frac{1}{T}\right)+2\left(\eta r\left(\alpha-\rho\frac{1}{T}\right)-b\left((\alpha+1)-\rho\frac{1}{T}\right)^2\right)\frac{1}{T} \\
&= \frac{\eta r}{T}\left(\alpha-\frac{\rho}{T}\right)+r\left(\alpha-\frac{\rho}{T}\right)-2b\left((\alpha+1)-\frac{\rho}{T}\right)^2\frac{1}{T}, \qquad \text{(A90)}
\end{aligned}$$

which is equivalent to (A86).

Third, we show that it does alter optimal solution for the payer problem to include the redundant constraint (A87). Note that the Nash equilibrium is bounded above by $\bar{\theta}$. If for a given value for η such that $\hat{\theta} \geq \bar{\theta}$, then the Nash equilibrium service level for hospitals should be equal to $\bar{\theta}$. On the other hand, checking the payer's objective function shows that the payer would incur lower costs when η takes a new value such that $\hat{\theta} = \bar{\theta}$. That is, optimal solutions for the payer problem are retained even if we impose the constraint $\hat{\theta} \leq \bar{\theta}$. Based on the single-crossing property, $\hat{\theta} \leq \bar{\theta}$ is equivalent to the following condition:

$$\begin{aligned}
0 &= \frac{\partial \pi_i^{(2)}}{\partial \theta_i}(\hat{\theta}, \hat{\theta}) \\
&\geq \frac{\partial \pi_i^{(2)}}{\partial \theta_i}(\bar{\theta}, \bar{\theta}) \\
&= r\left(1-\frac{\eta}{T}\right)\left(\alpha-\rho\bar{\theta}\right)+2\left(\eta r\left(\alpha-\rho\bar{\theta}\right)-b\left((\alpha+1)-\rho\bar{\theta}\right)^2\right)\bar{\theta} \\
&= \eta r\left(\alpha-\rho\bar{\theta}\right)\left(2\bar{\theta}-\frac{1}{T}\right)+r\left(\alpha-\rho\bar{\theta}\right)-2b\left((\alpha+1)-\rho\bar{\theta}\right)^2\bar{\theta}, \qquad \text{(A91)}
\end{aligned}$$

which is equivalent to (A87).

Fourth, in order to avoid an empty feasible region for the above equivalent optimization problem in Lemma A1, it is necessary to ensure that the right-hand-side of (A86) is less than or equal to the right-hand side of (A87). Assume $T \geq \frac{1}{2\theta^{\mathrm{NE}}(0)}$. Proposition 2 shows that $\theta^{\mathrm{NE}}(T,\eta)$ is monotone increasing in η. Define $\eta_{\frac{1}{T}}$ as the right-hand side of (A86) and $\eta_{\bar{\theta}}$ as the right-hand side of (A87). Then, compatibility between constraints (A86) and (A87) is equivalent to the following constraints:

$$\eta_{\frac{1}{T}} \leq \eta \leq \eta_{\bar{\theta}}. \qquad \text{(A92)}$$

Note that $\eta_{\frac{1}{T}}$ and $\eta_{\bar{\theta}}$ are derived from the best response equation $\frac{\partial \pi_i^{(2)}}{\partial \theta_i}(\theta,\theta)=0$ by assuming $\theta=\frac{1}{T}$ and $\theta=\bar{\theta}$, respectively. Since the Nash equilibrium is unique and symmetric, $\frac{\partial \pi_i^{(2)}}{\partial \theta_i}(\theta,\theta)=0$ has

a unique root that corresponds to the unique Nash equilibrium. Furthermore, this unique Nash equilibrium must satisfy $\theta^{\mathrm{NE}}(T,\eta_{\frac{1}{T}})=\hat{\theta}=\frac{1}{T}$ when $\eta=\eta_{\frac{1}{T}}$ because $\theta=\frac{1}{T}$ is a root of $\frac{\partial\pi_i^{(2)}}{\partial\theta_i}(\theta,\theta)=0$, and this unique Nash equilibrium must satisfy $\theta^{\mathrm{NE}}(T,\eta_{\bar{\theta}})=\hat{\theta}=\bar{\theta}$ when $\eta=\eta_{\bar{\theta}}$ because $\theta=\bar{\theta}$ is a root of $\frac{\partial\pi_i^{(2)}}{\partial\theta_i}(\theta,\theta)=0$. Suppose $\eta_{\frac{1}{T}}>\eta_{\bar{\theta}}$. Then, Proposition 2 shows that $\theta^{\mathrm{NE}}(T,\eta_{\frac{1}{T}})>\theta^{\mathrm{NE}}(T,\eta_{\bar{\theta}})$. Hence, $\frac{1}{T}>\bar{\theta}$, which is a contradiction. Hence, we have $\eta_{\frac{1}{T}}\le\eta_{\bar{\theta}}$ and constraints (A86) and (A87) are compatible. □

Proof of Proposition 9

(a) Assume $T\in[T_l,T_h]$. Then, we have $T\le T_h<\frac{1}{2\theta^{\mathrm{NE}}(0)}$. Corollary 1 shows that θ^{NE} is decreasing in η. Therefore, for any fixed T, there does not exist any η such that the corresponding unique Nash equilibrium $(\theta^{\mathrm{NE}},\theta^{\mathrm{NE}})$ for the hospital problem satisfies the service-level constraint $\theta^{\mathrm{NE}}\ge\frac{1}{T}$. Because $T_h<\frac{1}{2\theta^{\mathrm{NE}}(0)}$, there is no feasible solution for the payer problem.

(b) In order to prove the result, we decompose the payer problem into a nested optimization problem with two layers. In the inner layer, T is fixed and η is the decision variable, and in the outer layer, T is the only decision variable and η takes the optimal solution of the inner optimization problem when T is fixed. Once the inner optimization problem is solved, it is straightforward to solve the outer optimization problem. Hence, we focus on solving the inner optimization problem.

Assume $T\in[T_l,T_h]$. Then, we have $T\le T_l>\frac{1}{\theta^{\mathrm{NE}}(0)}$. Because of constraints (A86) and (A87), Lemma A1 shows that the unique Nash equilibrium $(\theta^{\mathrm{NE}},\theta^{\mathrm{NE}})$ for the symmetric duopoly satisfies $\theta^{\mathrm{NE}}=\hat{\theta}$, where $\hat{\theta}\in[\frac{1}{T},\bar{\theta}]$. This fact enables us to include the following redundant constraint

$$\frac{1}{T}\le\theta_i, \tag{A93}$$

for hospital i in the hospital problem. Recall that $\hat{\theta}$ satisfies the optimal response equation (64), which shows that $\hat{\theta}$ is differentiable in η and $\frac{\partial\theta^{\mathrm{NE}}}{\partial\eta}$ is well defined. Let us take the partial derivative of the objective function for the payer problem with regard to η:

$$\begin{aligned}\frac{\partial\Pi}{\partial\eta}&=2r\left(\alpha\frac{\partial\theta^{\mathrm{NE}}}{\partial\eta}-2\rho\theta^{\mathrm{NE}}\frac{\partial\theta^{\mathrm{NE}}}{\partial\eta}\right)\left[1+\eta\left(\theta^{\mathrm{NE}}-\frac{1}{T}\right)\right]\\&\quad+2r\left(\alpha\theta^{\mathrm{NE}}-\rho\theta^{\mathrm{NE}}\theta^{\mathrm{NE}}\right)\left[\theta^{\mathrm{NE}}-\frac{1}{T}+\eta\frac{\partial\theta^{\mathrm{NE}}}{\partial\eta}\right]\\&=2r\left(\alpha-2\rho\theta^{\mathrm{NE}}\right)\frac{\partial\theta^{\mathrm{NE}}}{\partial\eta}\left[1+\eta\left(\theta^{\mathrm{NE}}-\frac{1}{T}\right)\right]\\&\quad+2r\left(\alpha-\rho\theta^{\mathrm{NE}}\right)\theta^{\mathrm{NE}}\left[\theta^{\mathrm{NE}}-\frac{1}{T}+\eta\frac{\partial\theta^{\mathrm{NE}}}{\partial\eta}\right].\end{aligned} \tag{A94}$$

We claim that when $T\ge\frac{1}{2\theta^{\mathrm{NE}}(0)}$, $\frac{\partial\Pi}{\partial\eta}\ge0$ for the following arguments. First, when (A86) is satisfied, $\theta^{\mathrm{NE}}\ge\frac{1}{T}$, and when (A87) is satisfied, $\theta^{\mathrm{NE}}\le\bar{\theta}$. Second, $\alpha-2\rho\theta^{\mathrm{NE}}\ge\alpha-2\rho\bar{\theta}$ holds because of (62). Third, $\frac{\partial\theta^{\mathrm{NE}}}{\partial\eta}\ge0$ because Corollary 1 shows that when $T\ge\frac{1}{2\theta^{\mathrm{NE}}(0)}$, θ^{NE} is increasing in η. Therefore,

we have proved that for any fixed T, the objective function for the payer problem is increasing in η when (A86) and (A87) are satisfied. In other words, the optimal solution for the payer problem is to take the minimum value that satisfies all constraints.

We now show that constraint (A86) is redundant if $T > \frac{1}{\theta^{\mathrm{NE}}(0)}$ because the right-hand side of constraint (A86) is non-positive. First, $T > \frac{1}{\theta^{\mathrm{NE}}(0)}$ implies that $\theta^{\mathrm{NE}}(0) > \frac{1}{T}$ and $\hat{\theta}(0) > \frac{1}{T}$ because $\theta^{\mathrm{NE}}(0) = \min(\hat{\theta}(0), \bar{\theta})$. Second, the optimal response equation and single-crossing property show that

$$\begin{aligned} r\left(\alpha - \rho\hat{\theta}(0)\right) - 2b\left(\alpha + 1 - \rho\hat{\theta}(0)\right)^2 &= 0 \\ r\left(\alpha - \rho\frac{1}{T}\right) - 2b\left(\alpha + 1 - \rho\frac{1}{T}\right)^2 &> 0, \end{aligned} \tag{A95}$$

where the second inequality is equivalent to the following

$$0 > \frac{2b}{r}\frac{\left(\alpha + 1 - \rho\frac{1}{T}\right)^2}{\alpha + 1 - \rho\frac{1}{T}} - T. \tag{A96}$$

Therefore, the optimal solution for the payer problem is $\eta = 0$ and the optimal solution for the hospital problem is $(\theta^{\mathrm{NE}}(0), \theta^{\mathrm{NE}}(0))$.

Because for any $T \in [T_l, T_h]$, the optimal solution for the inner optimization problem is $\eta = 0$ and the optimal Nash equilibrium is $\theta^{\mathrm{NE}}(0)$. Thus, the objective function for the outer optimization problem does not depend on the decision variable T. Therefore, we conclude that the optimal solution for the payer problem is that $\eta = 0$ and T can take any value in $[T_l, T_h]$.

(c) Similarly to the proof for (b), we decompose the payer problem into the same nested optimization problem and focus on the inner optimization problem, for which T is a fixed parameter and η is the decision variable.

Assume $\frac{1}{2\theta^{\mathrm{NE}}(0)} \le T \le \frac{1}{\theta^{\mathrm{NE}}(0)}$. If $\eta^{\min}(T) > \eta^{\max}(T)$, then there is no feasible solution for the payer problem. Thus, we assume that $\eta^{\min}(T) \le \eta^{\max}(T)$, which implies that there is a feasible solution for the payer problem. First, $T \le \frac{1}{\theta^{\mathrm{NE}}(0)}$ implies that $\theta^{\mathrm{NE}}(0) \le \frac{1}{T}$ and $\hat{\theta}(0) \le \frac{1}{T}$ because $\theta^{\mathrm{NE}}(0) = \min(\hat{\theta}(0), \bar{\theta})$ and $\frac{1}{T} \le \bar{\theta}$. Second, the optimal response equation and single-crossing property show that

$$\begin{aligned} r\left(\alpha - \rho\hat{\theta}(0)\right) - 2b\left(\alpha + 1 - \rho\hat{\theta}(0)\right)^2 &= 0 \\ r\left(\alpha - \rho\frac{1}{T}\right) - 2b\left(\alpha + 1 - \rho\frac{1}{T}\right)^2 &\le 0, \end{aligned} \tag{A97}$$

where the second inequality is equivalent to the following

$$0 \le \frac{2b}{r}\frac{\left(\alpha + 1 - \rho\frac{1}{T}\right)^2}{\alpha + 1 - \rho\frac{1}{T}} - T. \tag{A98}$$

Thus the right-hand side of constraint (A86) is non-negative.

In (b) we have shown that the objective function for the inner optimization problem is increasing in η and that the optimal solution for the payer problem is achieved at the value for η such that (A86) becomes binding. Recall that Lemma A1 shows that $\theta^{\text{NE}} = \frac{1}{T}$ when (A86) is binding. Therefore, for any given value of T, the optimal value of η for the payer problem is to make $\theta^{\text{NE}} = \frac{1}{T}$ and to make (100) binding. Thus, we have proved

$$\eta = \frac{2b}{r} \frac{\left((\alpha+1) - \frac{\rho}{T}\right)^2}{\left(\alpha - \frac{\rho}{T}\right)} - T, \qquad \theta^{\text{NE}} = \frac{1}{T}. \tag{A99}$$

We now consider the outer optimization problem. For any fixed T, the results in (b) show that the optimal Nash service level is $\theta^{\text{NE}} = \frac{1}{T}$ and the optimal objective function value for the outer optimization problem is

$$\begin{aligned} \Pi(\eta(T), T) &= 2\left(\alpha\theta^{\text{NE}} - \rho\theta^{\text{NE}}\theta^{\text{NE}}\right) r \left[1 + \eta(T)\left(\theta^{\text{NE}} - \frac{1}{T}\right)\right] \\ &= 2r\left(\alpha - \rho\frac{1}{T}\right)\frac{1}{T}. \end{aligned} \tag{A100}$$

Taking the derivative of Π with respect to T, we obtain

$$\begin{aligned} \frac{\partial \Pi(\eta(T), T)}{\partial T} &= -2r\left(\alpha - 2\rho\frac{1}{T}\right)\frac{1}{T^2} \\ &< 0, \end{aligned} \tag{A101}$$

where the inequality follows from (62), $\bar{\theta} \geq \frac{1}{T}$ and $0 < \alpha - 2\rho\bar{\theta} \leq \alpha - 2\rho\frac{1}{T}$. Therefore, the objective function for the outer optimization problem is decreasing in T, where $\eta(T)$ is the optimal value for η for the payer problem when T is given and θ^{NE} is the unique Nash equilibrium for the hospital problem when both η and T are given. However, the payer problem may not have a feasible solution for some $T \in [T_l, T_h]$. Thus, it is straightforward to show the optimal value for T is T^*, where T^* is defined in (110). The desired results follow. □

Proof of Corollary 2

Following Proposition 9 (c), we have

$$\frac{\partial \eta}{\partial \alpha} = \frac{2b}{r}\left(1 - \frac{1}{\left(\alpha - \frac{\rho}{T}\right)^2}\right). \tag{A102}$$

This shows that the optimal value for the incentive parameter η is increasing in the demand-sensitivity parameter α if and only if $\alpha \geq \frac{\rho}{T} + 1$. Proposition 9 (b) shows that the Nash equilibrium service level remains $\frac{1}{T}$ even if α increases.

Assume $\alpha \geq \frac{\rho}{T} + 1$ ($\alpha \leq \frac{\rho}{T} + 1$) holds. The service level θ^{NE} remains $\frac{1}{T}$. The demand increases (decreases) because the demand is equal to $\alpha\theta^{\text{NE}} - \rho\theta^{\text{NE}}\theta^{\text{NE}}$. The cost for the payer increases

(decreases) following from the formula for the payer's objective function and that $\theta^{\mathrm{NE}} = \frac{1}{T}$. The cost for hospitals increases (decreases) following from the formula for the hospital's cost function and that $\theta^{\mathrm{NE}} = \frac{1}{T}$.

Following Proposition 9 (c), we have

$$\frac{\partial \eta}{\partial \rho} = \frac{2b}{rT}\left(-1 + \frac{1}{\left(\alpha - \frac{\rho}{T}\right)^2}\right). \tag{A103}$$

This shows that the optimal value for the incentive parameter η is decreasing in the demand-sensitivity parameter ρ if and only if $\alpha \geq \frac{\rho}{T} + 1$. Proposition 9 (c) shows that the Nash equilibrium service levels remain $\frac{1}{T}$ even if ρ increases.

Assume $\alpha \geq \frac{\rho}{T} + 1$ ($\alpha \leq \frac{\rho}{T} + 1$) holds. The service level θ^{NE} remains $\frac{1}{T}$. Demand decreases (increases) because demand is equal to $\alpha\theta^{\mathrm{NE}} - \rho\theta^{\mathrm{NE}}\theta^{\mathrm{NE}}$. The cost for the payer decreases (increases) following from the formula for the payer's objective function and that $\theta^{\mathrm{NE}} = \frac{1}{T}$. The cost for hospitals decreases (increases) following from the formula for the hospital's cost function and that $\theta^{\mathrm{NE}} = \frac{1}{T}$. □