

Pricing and Capacity Design for Profit-Driven and Welfare-Driven Healthcare Providers

Production and Operations Management
2024, Vol. 33(4) 1014–1030
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/10591478241238969
journals.sagepub.com/home/pao



Shengya Hua^{1,*} , Ying Lei^{2,*} and Xin Zhai³

Abstract

In choosing healthcare services, price and waiting time are two important factors that matter to patients. Price is set by healthcare providers, while waiting time is usually endogenously determined by patient choice and the healthcare provider's capacity investment. We study the pricing and capacity design for profit-driven hospitals offering two types of healthcare services—regular and premium—serving heterogeneous patients who choose from either of the two types of services. Patients make their choices based on both price and endogenously determined equilibrium waiting time. We then benchmark profit-driven hospitals to welfare-driven hospitals to reveal how the behaviors of for-profit hospitals deviate from the socially optimal outcomes in patients waiting, service capacity, and price. We find that fewer patients are treated by premium services in profit-driven hospitals, and thus profit-driven hospitals invest less in premium service capacity and charge a higher premium for premium services than welfare-driven hospitals. These inefficiency distortions exist in profit-driven hospitals primarily because they are incentivized to differentiate between the waiting times of the two types of service to induce patients to pay a higher premium for premium services. Our results also show that if the inefficiency in patient partition is corrected, profit-driven hospitals would choose the welfare-maximizing level of premium service capacity and thus achieve socially optimal results. However, we also find that regulations such as price ceiling and capacity regulation cannot fully correct the inefficiency in patient partition and capacity decisions of profit-driven hospitals. Lastly, the model is extended in several ways to ensure robustness.

Keywords

Pricing, capacity, healthcare, patient partition, policy intervention

Date received 25 November 2021; accepted 10 February 2024 after two revisions

Handling Editor: Sergei Savin

1 Introduction

Healthcare systems across the world face a range of challenges, including rising costs, aging populations, and a growing demand for services. Despite the efforts of policymakers, researchers, and practitioners to address these issues, there remains a shortage of resources such as doctors, nurses, and hospital beds. This shortage results in hospital overcrowding, leading to long waiting times for patients and increasing management burden for hospitals. One of the fundamental challenges in healthcare operations is providing affordable, inclusive, and prompt access to quality healthcare (Dai and Tayur, 2020). To make healthcare more affordable and inclusive, many countries provide public health insurance that grants people access to basic healthcare services without co-payments, but this further imposes a demand burden that leads to long waiting times for patients due to supply-side shortages.

Governments around the world have attempted to address the issue of long waiting times for healthcare services by increasing supply of resources and budgets. For example, the number of doctors per 1000 people in China increased from 2.2 in 2019 to 2.5 in 2021 (OECD, 2022). However, this approach has struggled to keep pace with the growing demand due to population aging in both developed and developing countries.

¹ School of Economics and Management, South China Normal University, Guangzhou, China

² Business Department, New York University Shanghai, Shanghai, China

³ Guanghua School of Management, Peking University, Beijing, China

*These two authors contributed equally to this work.

Corresponding author:

Xin Zhai, Guanghua School of Management, Peking University, No. 5, Yi He Yuan Road, Haidian District, Beijing, China.
Email: xinzhai@gsm.pku.edu.cn

Premium healthcare services coexist with basic/regular services but require higher co-payments or are less covered by health insurance. These premium services are typically staffed by experts and offer more flexible appointment times, including evening hours. Although premium services are priced higher than basic services due to the high cost of staffing and added convenience, the advantage is that patients have shorter waiting times and receive healthcare services from more senior and experienced doctors. Propper (1990) found that patients are willing to pay for a reduction in waiting time, and their willingness to pay increases with income. Therefore, this creates a two-tier healthcare system where some patients may choose premium service for its high treatment value while others may choose premium service simply to avoid long waiting times. Knowing that patients choose between regular and premium services based on treatment value, price, and waiting time of each service, how should a for-profit hospital that offers a two-tier healthcare system determine the price and capacity of the two types of services? Would a for-profit hospital intentionally create overcrowding in the regular service department in order to push patients to pay a higher price for premium services?

A large volume of empirical studies has compared the performance of hospitals with different ownership (i.e., for-profit hospitals vs. nonprofit hospitals) in terms of accessibility, efficiency, and quality of care. By summarizing those findings, Rosenau and Linder (2003) and Kruse et al. (2018) found in the United States and the European Union, respectively that for-profit hospitals do not perform as well as nonprofit or public hospitals, which contradicts the intuition of many health economists and policy makers. Despite the large volume of empirical evidence, an analytical framework is still lacking to understand the difference between the behaviors of for-profit hospitals and public hospitals and explain why for-profit hospitals are often outperformed by public hospitals. To fill the gap in the literature, we build an analytical model to study the capacity and pricing design of a two-tier healthcare system from two perspectives of hospitals: profit-maximizing versus welfare-maximizing, which correspond to the objectives of for-profit hospitals and nonprofit hospitals, respectively. In particular, our study aims to answer the following questions: How do for-profit hospitals and nonprofit hospitals differ in their decisions regarding patient partition between the two types of services (i.e., regular service and premium service) and managing capacity levels of the two types of services? What is the difference in outcomes or performance in terms of average quality of care, average waiting time, and service accessibility between these two scenarios (profit-maximizing and welfare-maximizing)? Can the profit-maximizing distortions in the behaviors of for-profit hospitals be corrected through policy interventions?

We model a hospital that offers both regular and premium services as a healthcare system with two types of servers: regular doctors and superior doctors. These two types of doctors differ in their skills and capabilities for providing

healthcare services. Specifically, superior doctors are more experienced in dealing with complicated healthcare needs, and thus the treatment values provided by superior doctors are perceived as higher by patients. In the context of healthcare services, patients' medical needs may require services with different levels of complexity. Patients with more complicated needs would value superior doctors even more than patients with simple needs. In addition, the two types of services may have different prices and different waiting times. Therefore, each patient's choice of healthcare service is jointly determined by the perceived value of treatment, price, and cost of waiting time of each type of service.

In the first part of this research, we start with a profit-maximizing scenario where the hospital chooses the price and capacity of regular and superior doctors to maximize its profit, knowing that price and capacity levels jointly determine patient choice with a rational expectation of equilibrium waiting time. A key novel feature of our model is that it accounts for the fact that patients evaluate quality and waiting time of a service differently by explicitly modeling healthcare service value for patients on these two dimensions. Moreover, the waiting time is endogenously determined by patient choice in the model. Since healthcare is related to public wellbeing, social welfare should be considered when analyzing optimal capacity and pricing decisions for healthcare systems. We then study a welfare-maximizing scenario representing socially optimal outcome (first-best). We benchmark the profit-maximizing scenario to the welfare-maximizing scenario to reveal how for-profit hospitals deviate from socially optimal outcomes in capacity, price, accessibility, quality of care, and patients waiting. We find that the welfare loss of a for-profit system is entirely due to the profit-maximizing distortion in patient partition. We then consider two types of policy interventions, price regulation and capacity regulation on for-profit hospitals, intending to reduce the welfare loss in a for-profit system. However, we find that these interventions cannot fully eliminate the welfare loss.

In the second part of this research, we extend our baseline model to a series of more realistic scenarios and demonstrate the robustness of our results. First, we assume full market coverage in the baseline model to focus on the main insights of the optimal pricing and capacity design of the two-tier system. By relaxing this assumption in the first extension model we show that our main results hold even if some patients with low willingness-to-pay can choose not to get treated. In the next extension model, we further relax the assumption in the baseline model that the capacity of regular doctors is given at the stage of decisions, and our results still hold. Finally, we consider a second dimension of patient heterogeneity in price sensitivity in addition to the dimension of complexity in healthcare needs in the last extension model.

This study makes several theoretical and practical contributions to the field. First, we build a theoretical model to study the optimal pricing and capacity design in a two-tier healthcare system, considering the effects of treatment value,

waiting time, consumer heterogeneity in both individual needs and price sensitivity. To the best of our knowledge, this is the first theoretical study of a two-tier service system that simultaneously examines quality differences on the supply side and consumer heterogeneity in multiple dimensions. Second, our analysis provides a robust theoretical framework for scenarios where for-profit hospitals employ price not only as a revenue tool but also as a means to manage waiting times, especially when patients care about both service quality and waiting time. This framework is general and applicable to other service contexts where both quality and time affect consumer choices. Third, by benchmarking the welfare-maximizing scenario as the socially optimal outcome, this study reveals how the behaviors of a for-profit hospital deviate from the socially optimal outcomes in service accessibility, quality of care, patients waiting, etc. Our analysis elucidates the mechanisms driving the prevailing yet counterintuitive empirical findings regarding hospital performance in the literature. These findings indicate that for-profit hospitals are often outperformed by nonprofit/public hospitals in service accessibility, average quality of care, and efficiency. Therefore, this study fills the gap in the literature and provides practical insights for policymakers in regulating the healthcare sector. Finally, our analysis captures the essential logic behind decision-making for both patients and hospitals. Specifically, we allow patients' waiting time to be endogenously determined within the model and find the Nash equilibrium of patient choice with rational expectation for equilibrium waiting time. We then investigate how the quality–waiting trade-off affects heterogeneous patients' choice and consequently affects hospital decisions on pricing and capacity.

The remainder of this paper is organized as follows. In Section 2, we review related literature. In Section 3, we present the baseline model and analyze hospital decisions in both profit-maximizing and welfare-maximizing scenarios. We also investigate the source of the inefficiency in the profit-maximizing scenario. In Section 4, we extend our baseline model in multiple directions and verify the robustness of our main results. We provide a conclusion with managerial implications in Section 5, along with directions for future research.

2 Literature Review

This study is framed as a queueing model, where price and capacity are two key measures to reduce congestions or long waiting times. Pricing and capacity decisions in queueing systems have long been studied. For example, Dewan and Mendelson (1990) consider the optimal pricing and capacity decisions to improve the efficiency of a service facility with nonlinear delay cost functions. Stidham (1992) investigates the stability and convergence of pricing and capacity decisions to a stationary policy in a queueing model. Maglaras et al. (2017) study the optimal pricing and capacity decisions in communication and information service systems. Interested readers

may refer to Stidham (2009) for a review of pricing and capacity decisions in queueing systems. However, all these studies focus on single queue systems, also known as one-tier service systems.

As we incorporate two types of services (regular service and premium service), our study is closely related to the literature on multiple types of servers, particularly two-tier service systems that include both superior servers and standard servers, such as hospitals, call centers, and online media platforms (Coban et al., 2019; Du et al., 2014; Hasija et al., 2005; Shumsky and Pinker, 2003). Such studies mainly examine two-level service systems that perform different types of tasks and investigate the assignment of tasks to lower- and higher-level servers (Saghafian et al., 2018; Shumsky and Pinker, 2003). Some studies explore both workload assignment and staffing problems in two-tier systems. For example, Hasija et al. (2005) study the scenario that all tasks first arrive at regular servers, where they are diagnosed and, if necessary, referred to superior servers. They use the square-root staffing rule to determine approximately optimal staffing levels and explore the optimal assignment rate based on task complexity. Lee et al. (2012) construct a more general assignment mechanism by relaxing the constraints on the treatment function in Hasija et al. (2005). Tarakci et al. (2009) explore optimal regular server investment and assignment policies for situations where the share of tasks and regular servers' treatment abilities are determined by the level of investment. Bimpikis and Markakis (2019) examine a case in which regular servers can only manage tasks that are below a certain level of complexity. However, all of the above-mentioned studies assign consumers to different servers and do not allow consumers to choose servers according to their requirements. In contrast, our model allows patients to choose from different types of servers in a profit-maximizing scenario and separates service quality from the value of a shorter waiting time. This enables us to study the interaction between the two dimensions of service value (i.e., higher quality of service vs. shorter waiting time) and investigate the optimal pricing and capacity decisions in a two-tier healthcare system.

It is widely known that in a congested service system, customers' self-interest results in a non-socially optimal equilibrium (Hassin and Haviv, 2003), meaning that a service must be appropriately priced to induce customers to behave in a socially optimal manner. Literature reviews by Hassin and Haviv (2003) and Hassin (2016) show that in a two-tier system, it is expected that customers with high delay costs pay a premium for quick service, while those with low delay costs prefer a free service that is not quick. However, this only holds true if the two servers offer the same service quality. In our study, we relax this restriction by explicitly modeling quality differences in regular and superior doctors. We also consider consumer heterogeneity in both service complexity and price sensitivity and investigate how these factors impact pricing and capacity decisions in a two-tier healthcare system.

A closely related work is Zhang and Yin (2021), where they consider a public service system consisting of two service providers with different service capacities and quality. Customers are delay sensitive in both channels and have strong preferences for toll service providers. The goal of the system is to induce customers to switch from the crowded toll channel to the less-congested free channel by pricing the toll service appropriately. Zhang and Yin (2021) use a queueing model to analyze customer choice and show that, under certain conditions, this two-tier system can reduce congestion and total social cost. Our study differs from Zhang and Yin (2021) in several ways. Firstly, we study a two-stage game where the healthcare facility determines the price and capacity of its two types of services in the first stage. In the second stage, patients choose which type of service (regular doctors or superior doctors) they wish to use in the profit-maximizing scenario. Secondly, we incorporate consumer heterogeneity in service complexity and allow consumers to choose doctors of different types based on their utility, which is a key feature of healthcare facilities. Thirdly, we relax the full-market coverage assumption, while Zhang and Yin (2021) assume all customers should be treated by either type of servers. Additionally, we incorporate consumer heterogeneity in price sensitivity and explore its impact on the hospital's capacity and pricing decisions in our model extension. Fourthly, we explicitly separate doctor service quality from the value of a short waiting time, allowing us to study the effects of service value on these two dimensions of patients' choice and on healthcare facilities' decisions.

3 The Baseline Model

We study a hospital that offers two types of doctor service: regular and superior. The services differ in their levels of service quality, with superior doctors provide at least the same as or strictly higher treatment value than regular doctors for all medical needs. Patients arrival follows a Poisson process with a mean arrival rate of λ . The patients are heterogeneous in the complexity levels of their medical needs. Let $k \sim U(0, 1)$ denote the complexity of a patient's medical need, which is assumed to follow a uniform distribution with the support of $[0, 1]$ (Shumsky and Pinker, 2003; Tarakci et al., 2009). We provide a full list of notations used in our baseline model in Appendix A.

A patient with medical needs of complexity k derives a total treatment value of $V + R(k)$ if treated by a regular doctor and a total treatment value of $V + \varepsilon R(k)$ with $\varepsilon > 1$ if treated by a superior doctor. The treatment value measures the increase in a patient's health due to a treatment-related reduction in the discomfort of an illness and is decomposed into two terms. The first term V is constant across patients and represents the base value of treatment. The second term $R(k)$ denotes the complexity-dependent treatment value. For tractability in the following analysis, we assume $R(k) = k$.

Mathematically, the total treatment value $V + \varepsilon k$ is a supermodular function of service quality and patient illness complexity, indicating that the difference between the treatment value provided by superior doctors and regular doctors increases with the complexity level k of a patient's medical needs.

The treatment capacities of the regular service and the superior service are denoted by μ_r and μ_s per unit of time, respectively. Following Hasija et al. (2005), we assume that μ_r and μ_s are independent of patient complexity level k for model tractability. Another justification for this assumption is that Lee et al. (2012) numerically show that, in a two-level service system with an M/M/N queue model setting, taking mean service time to be independent of complexity level does not significantly change the main results. As the total customer arrival rate is λ , we denote the effective arrival rates for the regular and superior services as λ_r and λ_s , respectively, with $\lambda_r + \lambda_s = \lambda$. Similar to Rajan et al. (2019) and Zhang and Yin (2021), we model each type of service as an M/M/1 queue and the expected waiting time of each service is given by

$$W_j = \frac{1}{\mu_j - \lambda_j}, \quad j = r, s. \quad (1)$$

Following Marchand (1974), we assume that the opportunity cost per unit of time, denoted as c_w , is constant, such that the total waiting cost is linear with respect to waiting time. Therefore, the waiting costs for a patient who chooses the superior doctor service or the regular doctor service are $c_w W_s$ or $c_w W_r$, respectively.

To focus on the main insights of the optimal design of the two-tier healthcare system using differential pricing and endogenous waiting time, we assume full-market coverage in the baseline model, as in many papers in the literature (e.g., Anderson and Renault, 2009; Jeon et al., 2004). Specifically, we assume that the base value V of health benefit from being treated is large enough that no patient chooses not to be treated for any illness. The full-market coverage assumption is not essential for our main results, and we will relax this assumption in the extension (Section 4.1). Meanwhile, there are two justifications in the context of healthcare services for this assumption. First, the non-treatment of any illness, even one with mild symptoms, can result in long-lasting adverse effects. Second, there is a global commitment to achieving universal health coverage, ensuring that everyone can access healthcare without being restricted by financial hardships (WHO, 2022). In many countries, the prices of basic medical treatments are subsidized and regulated to ensure that at least the basic needs of human well-being are met. Therefore, health benefits relative to the price of basic treatment are often considerable in reality.

3.1 Profit-Maximizing Scenario

In this section, we consider the profit-maximizing scenario in which healthcare services are provided by for-profit hospitals. The timing of the model is as follows: In the first stage,

a for-profit hospital sets the capacity and price for regular and superior doctor services. In the second stage, patients with heterogeneous medical needs arrive and choose between the two types of doctor services. At the end of the model, patients pay for the service they have chosen and receive the healthcare service.

We analyze the profit-maximizing model using backward induction. That is, we first present the patients' choice in the second stage, after the hospital has made pricing and capacity decisions for the two types of service in the first stage.

3.1.1 Patients. The patients' net utility from the service of a doctor comprises (i) the treatment value minus (ii) the price for the chosen type of doctor and (iii) the cost of waiting. Therefore, the net utility obtained by a patient with complexity k from a regular doctor is

$$U_r(k) = V + k - p_r - c_w W_r. \quad (2)$$

The net utility obtained by a patient with complexity k from a superior doctor is

$$U_s(k) = V + \varepsilon k - p_s - c_w W_s. \quad (3)$$

Patients with varying levels of illness complexity choose to receive treatment from either a superior doctor or a regular doctor according to the prices and expected waiting times of each option. The heterogeneity in patients' medical needs leads to a heterogeneity in their preference of higher service quality, resulting in a partitioning of patients into two groups based on their doctor choice. A patient with illness complexity k chooses a regular doctor if and only if $U_r(k) \geq U_s(k)$. First, we show that patients' choices can always be described by a threshold strategy, meaning that patients choose a superior doctor if their illness complexity exceeds a certain threshold and choose a regular doctor otherwise.

LEMMA 1. *Given any price and the expected waiting times for regular doctors and superior doctors, there exists a threshold $k^* \in [0, 1]$ such that all patients with illness complexity $k \leq k^*$ prefer regular doctors and those with illness complexity $k > k^*$ prefer superior doctors.*

According to Lemma 1, the equilibrium arrival rates for regular doctor service and superior doctor service can be expressed as

$$\lambda_r = \lambda k^*, \quad \lambda_s = \lambda(1 - k^*). \quad (4)$$

At equilibrium, the threshold k^* is determined jointly by the hospital's pricing of superior and regular doctor services and the waiting times for each service, which are endogenously generated by the model.

Similar to the models of adoption with network effects, in which agents play a simultaneous adoption game with rational expectations of the network size (e.g., Armstrong and

Wright, 2007; Church and Gandal, 1992), in the second stage of our model, patients play a simultaneous adoption game with a rational expectation of waiting times for regular and superior doctors. At equilibrium, the patient arrival rates for regular doctors and superior doctors are given by Equation (4). Then, the equilibrium waiting time in the queues for regular doctors and superior doctors is $W_r = \frac{1}{\mu_r - \lambda k^*}$ and $W_s = \frac{1}{\mu_s - \lambda(1 - k^*)}$, respectively. Given a rational expectation of equilibrium arrival rates, the marginal patient k^* who is indifferent between choosing a regular doctor and choosing a superior doctor is expressed as follows:

$$k^* - c_w \frac{1}{\mu_r - \lambda k^*} - p_r = \varepsilon k^* - c_w \frac{1}{\mu_s - \lambda(1 - k^*)} - p_s. \quad (5)$$

Rearranging Equation (5) to be $c_w \left(\frac{1}{\mu_s - \lambda(1 - k^*)} - \frac{1}{\mu_r - \lambda k^*} \right) - p_r = (\varepsilon - 1)k^* - p_s$, we can see that the left-hand side decreases in k^* and the right-hand side (RHS) increases in k^* . Therefore, it is easy to determine that there is a unique solution to Equation (5).

With endogenous waiting times and rational expectations of waiting times for patients' choice of doctor, the waiting cost can be interpreted as a *network externality* of one patient's choice on the other patients. Thus, it is negative for the patients who choose the same type of doctor and positive for the patients who choose a different type of doctor. This is a key point that differentiates our study from the studies (Guo and Zhang, 2013; Qian et al., 2017; Zhang and Yin, 2021) that usually take waiting time as exogenously given.

For the partitioning of patients into two groups (in terms of whether they choose a regular doctor or a superior doctor) to be a Nash equilibrium, it must be impossible for a patient to switch to another group and receive a higher utility. To see this, consider the marginal patients in the regular doctor group whose complexity is close to k^* : $k = k^* - \delta, \delta \rightarrow 0$. If these patients switch to the superior doctor group and thus cause a decrease in the partition threshold k^* , the waiting cost of the superior doctor group increases and that of the regular doctor group decreases, meaning that patients who switch will be worse off. For the same reason, the marginal patients in the superior doctor group (i.e., $k = k^* + \delta, \delta \rightarrow 0$) are also prevented from switching to the other group. The following lemma establishes the equilibrium of this second-stage patient problem.

LEMMA 2. *Given the prices and treatment capacity of the system, there exists a unique Nash equilibrium of patients' choice between a regular doctor and a superior doctor with rational expectations of waiting times. In this equilibrium, patients with illness complexity $k \leq k^*$ choose regular doctors and those with illness complexity $k > k^*$ choose superior doctors, where $k^* \in [0, 1]$ is determined by Equation (5).*

3.1.2 Hospital. In the first stage, the hospital optimizes its profit by choosing prices and capacity of the two types of doctor services, knowing how patients would choose under each combination of price and capacity. The hospital's profit consists of the payments received from patients minus the staffing costs of regular and superior doctors. To focus on the main analysis of the relative price and time design of the two-tier system, we assume that the regular service capacity (μ_r) is exogenously given when the hospital makes decisions in the baseline model. Regular services or basic medical treatments are usually subsidized, so the capacity is often limited in size due to budget reasons. Another justification is that the regular service capacity is part of the basic operation of the hospital while superior service capacity can be flexibly adjusted; therefore, the regular service capacity, as a long-term decision, can be viewed as given when we discuss the hospital's short-term decisions on superior service capacity and price. We relax this assumption of fixed regular capacity in Section 4.2. With μ_r fixed, without loss of generality, we normalize the staffing cost of regular doctors to zero. On the other hand, we assume superior doctors have a higher staffing cost $C(\mu_s) = \alpha\mu_s$, with $\alpha > 0$. Therefore, in sum, the hospital's profit can be written as:

$$\Pi = \lambda_r p_r + \lambda_s p_s - \alpha\mu_s, \quad (6)$$

where λ_r and λ_s are the arrival rates for regular and superior doctor services, respectively, as given in Equation (4).

Equation (5) shows that, with full-market coverage assumption, patients' decisions in the second stage are not affected by the level of p_r but instead are affected by the price difference between two types of services $p_s - p_r$. For simplicity, we assume that the price of regular service is exogenously given in

the baseline model and relax this assumption in the extension. This simplification does not affect our main results and is often justified in practice because the price of basic healthcare service is often subsidized or covered by public health insurance and is often regulated to ensure full coverage of basic health services. If we let $\Delta p = p_s - p_r$, the hospital's profit can be rewritten as $\Pi = \lambda p_r + \lambda(1 - k^*)\Delta p - \alpha\mu_s$. Therefore, when p_r is regulated, the hospital's pricing decision is equivalent to choosing Δp , the premium charged for the superior doctors.

Given a patient choice threshold of k^* , there is a one-to-one correspondence between the waiting time in superior doctor queue W_s and superior doctor capacity μ_s . That is, given the threshold k^* and the waiting time for superior service W_s , the superior doctor capacity is uniquely determined by $\mu_s = \frac{1}{W_s} + \lambda(1 - k^*)$. Therefore, the hospital's problem of choosing price and capacity is equivalent to choosing the premium Δp and the waiting time W_s for the superior doctor service. This means that we can transform the for-profit hospital's optimization problem from one consisting of price and capacity to one consisting of price premium and waiting time for the superior service, as follows:

$$\Pi = \max_{\Delta p, W_s} \lambda(1 - k^*)\Delta p - \alpha \left(\frac{1}{W_s} + \lambda(1 - k^*) \right), \quad (7)$$

where k^* is the equilibrium patient choice threshold as determined by

$$k^* - c_w \frac{1}{\mu_r - \lambda k^*} = \epsilon k^* - c_w W_s - \Delta p. \quad (8)$$

The optimal premium and waiting time decisions can be described by the following first-order conditions.

$$\begin{cases} \frac{\partial \Pi}{\partial \Delta p} = \underbrace{\lambda(1 - k^*)}_{\text{Higher margin (superior)}} - \underbrace{\lambda \Delta p \frac{\partial k^*}{\partial \Delta p}}_{\text{Lower volume (superior)}} + \underbrace{\alpha \lambda \frac{\partial k^*}{\partial \Delta p}}_{\text{Saved staffing cost}} = 0, \\ \frac{\partial \Pi}{\partial W_s} = \underbrace{-\lambda \Delta p \frac{\partial k^*}{\partial W_s}}_{\text{Lower volume (superior)}} + \underbrace{\alpha \left(\frac{1}{W_s^2} + \lambda \frac{\partial k^*}{\partial W_s} \right)}_{\text{Saved staffing cost}} = 0. \end{cases}$$

In the first-order condition for premium $\left(\frac{\partial \Pi}{\partial \Delta p} \right)$, when the premium for the superior doctor service increases by one unit, the first term represents the direct revenue gain from the higher premium paid by patients who choose the superior doctor service, the second term represents the revenue loss due to marginal patients changing from choosing the superior doctor

service to choosing the regular service, and the last term represents the savings in the cost of staffing superior doctors because of the reduced demand for the superior doctor service. In the first-order condition for the superior doctor service's waiting time $\left(\frac{\partial \Pi}{\partial W_s} \right)$, when the waiting time for the superior doctor service increases by one unit, the first term represents

the revenue loss from the marginal patients who change from choosing the superior doctor service to choosing the regular doctor service, and the second term is the saving in the cost of staffing superior doctors. It is easy to determine that there is a unique pair of optimal solutions $(\Delta p^{PM}, W_s^{PM})$ to the first-order conditions of the hospital's problem. The superscript PM denotes the optimal results of the profit-driven hospital.

Let k^{PM} denotes the patient choice threshold in equilibrium, that is, $k^{PM} = k^*(\Delta p^{PM}, W_s^{PM})$. The threshold is implicitly determined by Equation (8), this prevents us from obtaining a closed-form solution of k^{PM} .

LEMMA 3. *The hospital's profit first decreases with c_w when $W_r^{PM} \leq W_s^{PM}$, and then increases with c_w when $W_r^{PM} > W_s^{PM}$.*

The results of Lemma 3 are graphically illustrated in Figure 1. The effect of time value on the hospital profit is non-monotonic but directly follows from the hospital and patient decisions, as Figure 1 shows. When the time value is sufficiently small, the hospital chooses to make the superior doctor queue moving slower than the regular doctor queue ($W_r^{PM} \leq W_s^{PM}$), which initially results in some patients switching from the superior doctor service to the regular doctor service (i.e., k^{PM} increases) to save time on waiting as c_w increases. As a result, the hospital's profit decreases because fewer patients are paying for the superior doctor service and the increase in the superior doctor service premium Δp^{PM} is not sufficient to cover the loss of patients from the superior doctor service. On the other hand, when the time value gets larger, as illustrated in Figure 1, the hospital makes the superior queue moving faster than the regular queue ($W_r^{PM} > W_s^{PM}$) by employing sufficiently more superior doctors. More and more patients switch from regular doctor service to the superior doctor service (i.e., k^{PM} decreases), and each patient's willingness-to-pay for the superior service increases because of the increasing time value and shorter waiting time. As a result, the hospital's profit increases with c_w in this region. In the numerical studies, the value of V is normalized to 0 without loss of generality.

3.2 Welfare-Maximizing Scenario

In this section, we study the welfare-maximizing scenario that gives the first-best socially optimal outcomes. We then benchmark the profit-maximizing scenario to the first-best outcomes to reveal how the profit-maximizing outcomes deviate from their first-best levels in patient partition, capacity, price, patients waiting time, etc. By doing this, we can understand how a for-profit hospital operates a healthcare system differently from a nonprofit or public hospital that aims to improve social wellbeing, and then we can look into policy interventions that may influence or regulate the organization to achieve socially optimal outcomes.

Social welfare comprises patients' health benefits from treatment minus the total waiting time and staffing costs of a

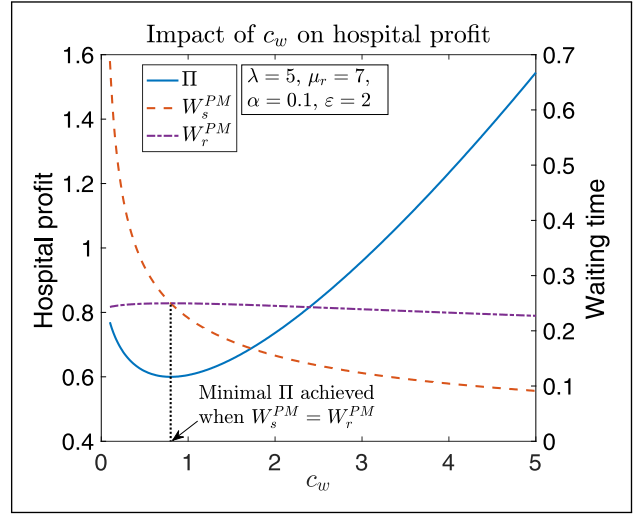


Figure 1. Effects of time value on hospital profit.

health system. Prices for services are simply transfers between the patients and the hospital and do not affect the total welfare. Therefore, the welfare-driven hospital directly chooses the patient partition threshold k^* and superior doctor service capacity μ_s to maximize

$$S = \max_{k^*, \mu_s} \lambda \int_0^{k^*} (V + k) dk + \lambda \int_{k^*}^1 (V + \epsilon k) dk - \lambda k^* c_w W_r - \lambda (1 - k^*) c_w W_s - \alpha \mu_s, \quad (9)$$

where W_r and W_s are determined according to Equation (1). The staffing costs of doctors in a competitive labor market are equal to the opportunity costs of their time spent providing service, which are the capacity costs of the healthcare system. In summary, the social benefits of healthcare services are the total treatment value received by all patients, while the social costs consist of the deadweight loss from the waiting time and the cost of building capacity. Again, the capacity and staffing costs of the regular doctor service are exogenous and suppressed in Equation (9). A welfare-driven hospital determines the patient partition threshold k^* and the superior doctor service capacity μ_s by balancing the average quality of service that patients receive and the congestion pressure on capacity of the healthcare system.

For a welfare-driven hospital, given a patient partition threshold of k^* , the capacity of superior doctor service μ_s has a one-to-one correspondence with the waiting time for superior doctor service W_s , that is, $\mu_s = \frac{1}{W_s} + \lambda(1 - k^*)$. Therefore, similar to the profit-maximizing scenario, we transform the welfare-driven hospital's problem of choosing the patient partition threshold and capacity (Equation 9) to the following problem of choosing the patient partition threshold and

waiting time for the superior doctor service:

$$\max_{k^*, W_s} \lambda \int_0^{k^*} (V + k) dk + \lambda \int_{k^*}^1 (V + \varepsilon k) dk - \lambda k^* c_w W_r - \lambda(1 - k^*) c_w W_s - \alpha \left(\frac{1}{W_s} + \lambda(1 - k^*) \right). \quad (10)$$

The optimal decisions of patient partition threshold and waiting time for the welfare-driven hospital can be described by the following first-order conditions:

$$\begin{cases} \frac{\partial S}{\partial k^*} = \underbrace{-\lambda(\varepsilon - 1)k^*}_{\text{Less superior treatment}} - \underbrace{\frac{\lambda c_w \mu_r}{(\mu_r - \lambda k^*)^2}}_{\text{Longer regular queue}} + \underbrace{\frac{\lambda c_w W_s}{\mu_r}}_{\text{Shorter superior queue}} + \underbrace{\lambda \alpha}_{\text{Saved staffing cost}} = 0, \\ \frac{\partial S}{\partial W_s} = \underbrace{-\lambda c_w(1 - k^*)}_{\text{Longer superior waiting time}} + \underbrace{\frac{1}{W_s^2} \alpha}_{\text{Saved staffing cost}} = 0. \end{cases}$$

Let k^{WM} and W_s^{WM} denote the first-best patient partition threshold and waiting time in the welfare-maximizing scenario, and let $\mu_s^{WM} = \frac{1}{W_s^{WM}} + \lambda(1 - k^{WM})$ denote the first-best capacity of superior doctors. Based on the first-best outcomes, we can get the next lemma.

LEMMA 4. *Social welfare strictly decreases with patients' time sensitivity c_w .*

We use a graphical decomposition in Figure 2 to show the underlying rationale of this result. The treatment value increases slightly with c_w when c_w is small and then remains constant as c_w increases. This is because when c_w is large enough, the hospital directs more patients to superior doctor services (k^{WM} decreases) and the total treatment value for society increases until all of the patients are directed to the superior doctor service ($k^{WM} = 0$), after which the treatment value can no longer be improved. The decrease in total welfare is caused by the welfare loss due to patients' increasing time sensitivity and the increased staffing cost associated with the reduced waiting time. Figure 2 also shows that the total waiting cost for society increases nonlinearly with c_w in a concave form, indicating that the total waiting time decreases as c_w increases. The reduction in the total waiting time is due to the increased capacity investment.

3.3 Comparison of the Two Scenarios

If the equilibrium patient partition in the profit-maximizing scenario is different from the equilibrium patient partition in the welfare-maximizing scenario, this means that for-profit hospitals reduce the efficiency of the healthcare system and cause welfare loss. In Section 3.4, we also discuss a few policy interventions to correct such inefficiencies.

3.3.1 Patient Partition and Capacity Decisions. Although the equilibrium threshold of patient choice k^{PM} on the

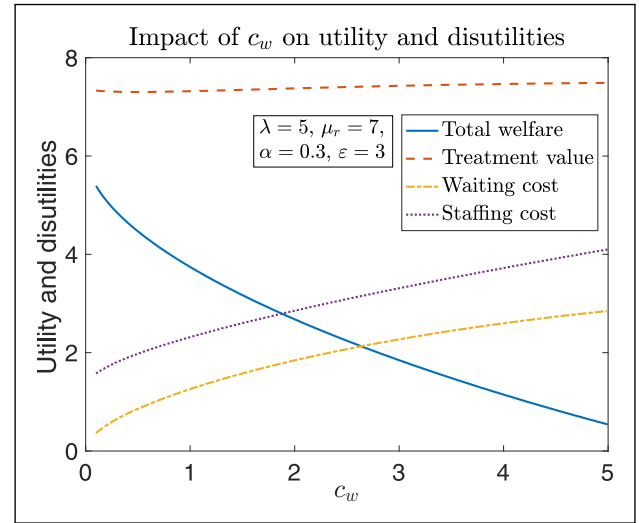


Figure 2. Effect of time value on social welfare with decomposition.

profit-driven hospital is implicitly determined by Equation (8) at equilibrium and does not have a closed-form solution, we can compare the optimal decisions in the two scenarios by comparing their first-order conditions. To facilitate comparison, we plug the patient choice equilibrium into the profit-maximizing problem in Equation (7) and transform the first-order condition on premium Δp to a first-order condition on the patient partition threshold k^* . This gives the following relationship between the profit-maximizing threshold and the welfare-maximizing threshold conditions:

$$\frac{\partial \Pi}{\partial k^*} - \frac{\partial S}{\partial k^*} = \underbrace{\lambda k^* c_w \frac{\partial W_r}{\partial k^*}}_{\text{Extra waiting cost (regular)}} + \underbrace{\lambda(1 - k^*) \frac{\partial \Delta p}{\partial k^*}}_{\text{Revenue cost of patient partition}}, \quad (11)$$

where $\frac{\partial \Delta p}{\partial k^*}$ is the change in the premium required to achieve a unit change in the patient partition threshold k^* . Equation (11) shows that the difference between a welfare-maximizing perspective and a profit-maximizing perspective in patient partition threshold is that the profit-maximizing perspective does not internalize the patients' welfare loss from the waiting cost (the first term), while the welfare-maximizing perspective is not concerned about the revenue loss from the price change used to adjust the patient partition threshold (the second term). The results of the comparison of the first-order conditions in the profit-maximizing scenario with those in the welfare-maximizing scenario are presented in the following lemma.

LEMMA 5. *For any patient partition threshold k^* , it is always the case that $\frac{\partial \Pi}{\partial k^*} > \frac{\partial S}{\partial k^*}$ and $\frac{\partial \Pi}{\partial W_s} = \frac{\partial S}{\partial W_s}$.*

Lemma 5 shows that directing the marginal patients from the superior doctor service to the regular doctor service (i.e., a marginal increase in k^*) has a greater effect on profit than on social welfare. Therefore, the optimal patient partition threshold in the welfare-maximizing scenario, denoted by k^{WM} , which satisfies the first-order condition $\frac{\partial S}{\partial k^*} = 0$, is different from the equilibrium patient-choice threshold in the profit-maximizing scenario, denoted by k^{PM} , which must satisfy the first-order condition $\frac{\partial \Pi}{\partial k^*} = 0$.

Interestingly, Lemma 5 also shows that profit and social welfare obtain the same benefit from a shorter waiting time for the superior doctor service. Therefore, given the same patient partition threshold k^* , the hospital will choose the same waiting time for the superior doctor service and the same superior doctor service capacity in both profit-maximizing and welfare-maximizing scenarios. Thus, if the patient partition threshold is socially optimal, a profit-driven decision on capacity is the same as the socially optimal capacity. We formalize this conclusion in the next proposition.

PROPOSITION 1. *The inefficiency in the profit-maximizing scenario is entirely due to the distortion in the patient partition threshold k^* .*

The rationale for Proposition 1 is simple: the patient welfare loss due to the waiting in the superior doctor queue is incorporated into the premium that the patients are willing to pay for the superior doctor service. Therefore, the objective of optimizing waiting time for the superior doctor service in the profit-maximizing scenario is aligned with the welfare-maximizing scenario.

Next, we compare the major results of the profit-maximizing scenario with those of the welfare-maximizing scenario.

PROPOSITION 2. *Let (k^{PM}, μ_s^{PM}) denote the threshold of patient partition and the optimal capacity of superior doctor service in the profit-maximizing scenario, and let (k^{WM}, μ_s^{WM})*

denote the threshold of patient partition and the optimal capacity of superior doctor service in the welfare-maximizing scenario. Then we have that

- (i) *the profit-maximizing scenario directs fewer patients to the superior doctor service than the socially optimal number of patients, or $k^{PM} > k^{WM}$; and*
- (ii) *the profit-maximizing scenario has fewer superior doctors than the socially optimal superior doctor capacity, or $\mu_s^{PM} < \mu_s^{WM}$.*

When the waiting time for the superior doctor service is short, patients perceive it as another dimension of high value of this service. Therefore, a for-profit hospital has an incentive to reduce the queue length of the superior doctor service. This can be done in two ways. First, the hospital can direct more patients to the regular doctor service. This means that there are fewer patients for the superior doctor service, but they are willing to pay a premium because the shorter waiting time improves the total value. This is a *volume versus margin* trade-off for the hospital. Second, the hospital can increase the capacity of the superior doctor service by hiring more superior doctors, albeit at high cost. This is a *value versus cost* trade-off for the hospital. The ultimate optimal premium and capacity decisions of the for-profit hospital (and consequently the patient partition threshold) are obtained by simultaneously balancing these two trade-offs. However, a nonprofit hospital does not face the volume versus margin trade-off as it is not concerned with prices or profit margins. Thus, the optimal decisions for the welfare-driven hospital are the results of only the value versus cost trade-off. Therefore, Proposition 2 shows that fewer patients are directed to the superior doctor service by the for-profit hospital than by the nonprofit hospital because the latter has no incentive to increase the premium charged for the superior doctor service by keeping the superior doctor queue short. Consequently, combined with the results of Proposition 1, this shows that nonprofit hospitals that maximize social welfare invest more in superior doctor service capacity than for-profit hospitals simply because of the higher demand for the superior doctor service on the former hospitals than on the latter hospitals. Consistent with the findings summarized in Rosenau and Linder (2003) and Kruse et al. (2018), this implies that the average quality of healthcare service is lower in for-profit hospitals compared to that in nonprofit hospitals, since the former hospitals invest less in superior doctor service capacity and direct fewer patients to the superior doctor service.

Proposition 1 shows that the deviation of the profit-maximizing patient partition threshold from the first-best socially optimal threshold is the essential distortion under the profit-maximizing objective. Therefore, it is meaningful to study how this distortion changes with the model parameters. Next, we numerically investigate the change in the profit-driven distortion on patient partition threshold ($k^{PM} - k^{WM}$) as patients' time value c_w varies.

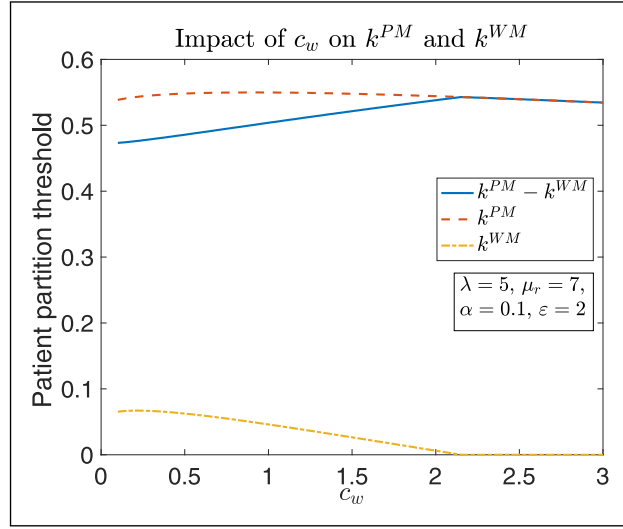


Figure 3. Sensitivity analysis of the distortion in patient partition threshold.

Figure 3 shows that when patients' time value increases (i.e., as c_w becomes larger), both k^{PM} and k^{WM} decrease, but k^{WM} decreases faster, and the difference between the two (i.e., the inefficiency or the distortion) continues increasing until k^{WM} becomes 0. The rationale for this is given by Equation (11). As patients care increasingly about time, the welfare loss caused by waiting increases, making the first term on the RHS of Equation (11) more negative, and the patient partition threshold becomes less sensitive to price, making the second term on the RHS of Equation (11) less positive. Consequently, the increase in c_w increases the misalignment in the optimized patient partition conditions between the profit-maximizing scenario and the welfare-maximizing scenario, thus increasing the profit-driven distortion in k^* .

3.3.2 Social Welfare Loss Due to Waiting. In this section, we analyze the difference in the equilibrium waiting time between the profit-maximizing scenario and the welfare-maximizing scenario. Because waiting costs are incurred by the patients and are a deadweight loss to society, we study the total deadweight loss from waiting in both scenarios.

PROPOSITION 3. *The equilibrium waiting time of both regular and superior doctor services are longer in the profit-maximizing scenario than in the welfare-maximizing scenario. That is, $W_j^{PM} > W_j^{WM}$, $j = r, s$.*

Proposition 3 concludes that the equilibrium waiting times of both types of services are longer in the profit-maximizing scenario, indicating that the overall level of timely access to healthcare services is lower in for-profit hospitals than in non-profit hospitals. Consistent with the findings summarized in Kruse et al. (2018), the results in Proposition 3 imply that non-profit hospitals outperform for-profit hospitals in the efficiency

of service reflected in waiting times. Consequently, since c_w denotes the opportunity cost per unit of time of a patient, Proposition 3 also concludes that, at equilibrium, the per capita waiting cost in equilibrium of both regular and superior doctor services are greater in the profit-maximizing scenario than in the welfare-maximizing scenario, that is $c_w W_j^{PM} > c_w W_j^{WM}$, $j = r, s$.

Although the per capita waiting cost in both the regular and superior doctor queues is greater in the profit-maximizing scenario, the total waiting cost in the profit-maximizing scenario is not necessarily higher than in the welfare-maximizing scenario, because a smaller group of patients is directed to the superior doctor service in the profit-maximizing scenario than in the welfare-maximizing scenario ($1 - k^{PM} < 1 - k^{WM}$). Let the total waiting cost of all patients who choose service type j in the profit-maximizing case be denoted by WC_j^{PM} , where $j = r, s$, and the total waiting cost of all patients who choose service type j in the welfare-maximizing case be denoted by WC_j^{WM} , where $j = r, s$. That is, $WC_j^i = \lambda k^i c_w W_r^i$ ($i = PM, WM$) and $WC_s^i = \lambda(1 - k^i) c_w W_s^i$ ($i = PM, WM$). This leads to the next proposition on the comparison between the total waiting cost of the regular doctor queue and the superior doctor queue in both scenarios.

COROLLARY 1. *The total waiting cost of all of the patients who choose the regular doctor service is higher in the profit-maximizing scenario than in the welfare-maximizing scenario, while the total waiting cost of all of the patients who choose the superior doctor service is lower in the profit-maximizing scenario than in the welfare-maximizing scenario. That is, $WC_r^{PM} > WC_r^{WM}$, and $WC_s^{PM} < WC_s^{WM}$.*

The total waiting cost incurred by a group of patients is jointly determined by the size of the group and the average waiting time of each patient. Recall from Proposition 2 that in the profit-maximizing scenario, more patients are directed to the regular doctor service than in the welfare-maximizing scenario, and recall from Proposition 3 that the average waiting times of the regular and superior doctor queues are both longer in the profit-maximizing scenario than in the welfare-maximizing scenario. It follows that the total waiting cost of the group of patients receiving the regular doctor service is higher in the profit-maximizing scenario (WC_r^{PM}) than in the welfare-maximizing scenario (WC_r^{WM}). Corollary 1 also shows that even though the individual waiting time in the superior doctor queue is longer in the profit-maximizing scenario than in the welfare-maximizing scenario, the number of patients in the superior doctor queue is reduced so much that the total waiting cost of the group waiting for superior doctor service is lower in the profit-maximizing scenario (WC_s^{PM}) than in the welfare-maximizing scenario (WC_s^{WM}). This profit-driven reduction in the total waiting cost of the patients treated by superior doctors, compared to the first-best level, is not due to a fast-moving queue for the superior doctor service.

Rather, it is due to a much shorter queue for the superior doctor service in the profit-maximizing scenario than in the welfare-maximizing scenario, as a for-profit hospital is incentivized to save on the staffing cost and to differentiate the waiting times between the regular doctor and the superior doctor queues to motivate patients to pay for the superior doctor service.

3.4 Policy Interventions

3.4.1 Price Regulation. From previous analysis, we can see that profit-driven hospitals tend to direct fewer patients to the superior service and charge a higher price for the superior service than the socially optimal level. In this section, we analyze policy intervention in the form of price regulation to see if this corrects the profit-driven inefficiency in the healthcare system. Price ceiling is a common policy tool used to ensure the basic wellbeing of a large population. In various regulated industries, such as electric power, governments establish a maximum price that can be charged for certain goods or services to prevent monopolies from excessively raising prices. In the healthcare sector, examples of price regulations also exist. For instance, the National Pharmaceutical Pricing Authority in India has implemented price caps on various medical procedures and treatments, particularly in private hospitals, to prevent overcharging and make healthcare more affordable (NIC, 2019). In Australia, the government sets a price ceiling for many medical services via the Medicare Benefits Schedule. Providers who choose to “bulk bill” agree to charge the government-set price and not to charge the patient any additional fees (Barber et al., 2019).

In this section, we consider a price ceiling at the socially optimal level (the first-best price), defined as follows. If we allow the welfare-driven hospital to induce socially optimal behaviors in patients through pricing, instead of through patient directing, we refer to the price (premium) that induces first-best outcomes as the first-best price Δp^{WM} . Given the first-best levels of patient partition k^{WM} and waiting time of superior doctor service W_s^{WM} , we calculate the first-best price of superior doctor service Δp^{WM} as follows:

$$\Delta p^{WM} = (\varepsilon - 1)k^{WM} + \frac{c_w}{\mu_r - \lambda k^{WM}} - c_w W_s^{WM}. \quad (12)$$

By direct comparison, it is evident that a for-profit hospital would charge a higher premium for the superior doctor service than the first-best level, that is, $\Delta p^{PM} > \Delta p^{WM}$. In the following, we analyze the decisions in the profit-maximizing scenario if the premium for the superior doctor service is regulated such that it cannot be higher than the socially optimal level, that is, $\Delta p \leq \Delta p^{WM}$, where Δp^{WM} is given in Equation (12).

First, we can see that the price regulation is binding (i.e., $\Delta p = \Delta p^{WM}$) at equilibrium. The reason is the follows. From the previous analysis of the profit-maximizing scenario, we know that profit is convex with respect to the superior doctor premium Δp . Suppose that under the price cap, the hospital's

profit-maximizing strategies for its premium and superior doctor service capacity are $(\Delta p', \mu_s')$, such that $\Delta p'$ is strictly lower than the price cap. This would imply that the hospital's profit has a local maximum at $(\Delta p', \mu_s')$, which contradicts the fact that profit is strictly convex in both Δp and μ_s and that the optimal unrestricted premium of superior service in the profit-maximizing scenario is greater than the first-best level (i.e., $\Delta p^{PM} > \Delta p^{WM}$). Therefore, increasing the premium of superior service from $\Delta p'$ to the price cap Δp^{WM} would increase the profit, and thus, $(\Delta p', \mu_s')$ cannot be a local maximum.

Then the profit-maximizing scenario under price regulation can be written as follows:

$$\max_{\mu_s} \lambda(1 - k^*)\Delta p^{WM} - \alpha\mu_s, \quad (13)$$

where k^* is determined by the patients' adoption equilibrium according to $k^* - c_w \frac{1}{\mu_r - \lambda k^*} = \varepsilon k^* - c_w \frac{1}{\mu_s - \lambda(1 - k^*)} - \Delta p^{WM}$. The hospital's optimal decisions under price regulation are summarized in the next proposition.

PROPOSITION 4. *When the premium for the superior doctor service is capped such that it is no higher than the welfare-maximizing premium Δp^{WM} , the profit-driven hospital charges the welfare-maximizing premium ($\Delta p = \Delta p^{WM}$) but continues to direct more patients to the regular doctor service ($k^* > k^{WM}$) and hire less superior doctors ($\mu_s < \mu_s^{WM}$) than the social optimal amount. This makes the waiting time in the superior doctor queue on the profit-driven hospital longer than that on the welfare-driven hospital ($W_s > W_s^{WM}$).*

Proposition 4 shows that even if the premium for the superior doctor service is regulated by applying a socially optimal price cap, a for-profit hospital will not choose the first-best staffing level, and thus, the inefficiency distortion in the system persists. In particular, the hospital charges the socially optimal price but invests less in the capacity for the superior doctor service, which makes the waiting time for superior doctor service longer than the socially optimal waiting time. Therefore, some patients who would have chosen the superior doctor service under the welfare-maximizing scenario instead choose the regular doctor service to avoid a long queue. This underinvestment in superior doctor service capacity results from the fact that for-profit hospitals do not internalize the loss in patient welfare resulting from the increased waiting when making their capacity decision.

3.4.2 Capacity Regulation. In this section, we consider another policy intervention: capacity regulation. Capacity regulations are common in industries facing an undersupply problem, such as public goods and the superior service in our model. When for-profit providers of certain resources or services tend to supply less than the socially optimal level, governments often mandate a minimum level of resources or services to meet public demand. In the healthcare sector,

for example, countries like the United States (Roberts, 2023) and Australia (ACT, 2023) have regulations requiring a minimum nurse-to-patient ratio in hospitals and nursing homes to ensure quality care and patient safety.

In the following, we analyze the profit-maximizing scenario if the superior doctor service capacity is regulated such that it cannot be lower than the socially optimal level, that is, $\mu_s \geq \mu_s^{WM}$. Using similar arguments as in the price regulation analysis, we show that capacity regulation is also binding (i.e., $\mu_s = \mu_s^{WM}$) at equilibrium as otherwise there would be a local maximum with $\mu'_s > \mu_s^{WM}$. However, since hospital profit is strictly convex in both Δp and μ_s and the optimal capacity at the global maximum is lower than the welfare-maximizing capacity (i.e., $\mu_s^{PM} < \mu_s^{WM}$), increasing the superior capacity from μ'_s to the binding level μ_s^{WM} would further increase the profit and leads to a contradiction. Therefore, such local maximum does not exist, and the capacity regulation will be binding.

The profit-maximizing scenario under capacity regulation can be written as follows:

$$\max_{\Delta p} \lambda(1 - k^*)\Delta p - \alpha\mu_s^{WM}, \quad (14)$$

where k^* is determined by patients' adoption equilibrium according to $k^* - c_w \frac{1}{\mu_r - \lambda k^*} = \epsilon k^* - c_w \frac{1}{\mu_s^{WM} - \lambda(1 - k^*)} - \Delta p$. The hospital's optimal decisions under capacity regulation are summarized in the next proposition.

PROPOSITION 5. *When the superior doctor service capacity is required to be no less than the welfare-maximizing superior doctor service capacity of μ_s^{WM} , the profit-driven hospital sets the superior doctor service capacity at the welfare-maximizing level ($\mu_s = \mu_s^{WM}$) but charges a higher premium for the superior doctor service than the welfare-maximizing premium ($\Delta p > \Delta p^{WM}$). This directs more patients to the regular doctor service ($k^* > k^{WM}$) than the welfare-maximizing level.*

Proposition 5 shows that capacity regulation cannot correct the inefficiency distortion in a for-profit hospital either. In particular, with capacity regulation, the profit-driven hospital staffs exactly at the socially optimal level but charges a higher premium for the superior doctor service than the welfare-driven hospital. With such a high premium for superior service, some patients with a medium level of illness complexity become discouraged and choose the regular doctor service instead, while patients with high levels of illness complexity remain with the superior doctor service and are willing to pay a high premium for service because their waiting time is short. The intuition underlying this result is that a short waiting time for the superior doctor service is more profitable for the for-profit hospital than a short waiting time for the regular doctor service, while a short waiting time for either service leads to the same welfare-loss for a welfare-driven hospital. Therefore, when capacity is regulated at the socially optimal level,

compared with a welfare-driven hospital, the profit-driven hospital has an additional incentive to reduce the waiting time for the superior doctor service.

4 Model Extensions

4.1 Flexible Market Participation

In the baseline model, we assume full-market coverage to focus on the main insights into the optimal pricing and capacity design of the two-tier healthcare system and the factors causing inefficiencies in the profit-maximizing scenario. In reality, patients may opt not to seek treatment at the hospital if their conditions are simple, and healthcare services are costly or crowded. In this section, we relax the full-market coverage assumption to better reflect reality and assess the robustness of our main findings on the distortion in patient partition thresholds and waiting times of profit-driven hospitals.

Denote k_r and k_s ($k_s \geq k_r$) as the thresholds of choosing regular service and superior service, respectively. In particular, the two thresholds are such that $U_r(k_r) = 0$ and $U_r(k_s) = U_s(k_s)$. Following a similar proof for Lemma 1, it is straightforward to verify that under flexible market participation, only patients with complexity $k \geq k_r$ choose to receive healthcare services. Additionally, patients with complexity $k \in [k_r, k_s]$ choose to be treated by regular doctors and patients with complexity $k > k_s$ choose superior doctors. We derive the condition for full-market coverage in the next lemma.

LEMMA 6. *All arrival patients choose to receive treatment in equilibrium (i.e., the market is endogenously “covered”) when*

$$\epsilon < \frac{2c_w^2 + \lambda V^3 [\lambda(V - \alpha) - \mu_r] - c_w V \left[2\mu_r - \lambda(2V - \alpha + 1) + \sqrt{\frac{\alpha V \lambda^2 (c_w + \lambda V^2)^2}{c_w(c_w - \mu_r V + V \lambda)}} \right]}{(c_w + \lambda V^2)[2c_w + V(\lambda - 2\mu_r)]}.$$

Lemma 6 suggests that in situations where the superior services do bit offer significantly greater value compared to regular services, the market will naturally achieve full “coverage.” In such cases, the use of the full-market coverage assumption, as employed in our baseline model, does not result in a loss of generality.

With flexible market participation, the inefficiency in the profit-maximizing scenario is entirely due to the distortion in patient partition thresholds k_r and k_s (a detailed proof is provided in the online appendix). This result confirms the robustness of Proposition 1 in the baseline model. The rationale is that flexible market participation does not change the fact that patient welfare loss due to waiting in the superior doctor queue is incorporated into the premium that the patient is willing to pay.

In what follows, we further explore the robustness of the comparison results between profit-maximizing and welfare-maximizing scenarios in the setting with flexible market participation.

PROPOSITION 6. Let $(k_r^{PM}, k_s^{PM}, \mu_s^{PM})$ denote the equilibrium complexity thresholds of patient partition and the optimal capacity of superior doctor service in the profit-maximizing scenario, and let $(k_r^{WM}, k_s^{WM}, \mu_s^{WM})$ denote the equilibrium complexity thresholds of patient partition and the optimal capacity of superior doctor service in the welfare-maximizing scenario. Then we have that

- (i) fewer patients choose to get treatment in the profit-maximizing scenario, that is, $k_r^{PM} > k_r^{WM}$;
- (ii) the profit-maximizing scenario directs fewer patients to the superior doctor service and more patients to the regular doctor service than the socially optimal number of patients, that is, $k_s^{PM} > k_s^{WM}$, and $k_s^{PM} - k_r^{PM} > k_s^{WM} - k_r^{WM}$;
- (iii) the profit-maximizing scenario has fewer superior doctors than the socially optimal superior doctor capacity, or $\mu_s^{PM} < \mu_s^{WM}$.

With flexible market participation, result (i) in Proposition 6 implies that the accessibility of healthcare service is lower in the profit-maximizing scenario than in the welfare-maximizing scenario, which is consistent with the findings summarized in Rosenau and Linder (2003) and Kruse et al. (2018) that for-profit hospitals usually have lower accessibility than nonprofit hospitals. Results (ii) and (iii) in Proposition 6 reaffirm the robustness of our findings regarding how a for-profit hospital's optimal patient partition and superior capacity decisions differ from the first-best welfare-maximizing levels (Proposition 2). In line with the baseline model, we observe that a for-profit hospital directs fewer patients to the superior service than the first-best level ($k_s^{PM} > k_s^{WM}$), while directing more patients to the regular service ($k_s^{PM} - k_r^{PM} > k_s^{WM} - k_r^{WM}$). The robustness of these results in the setting of flexible market participation underscores that strategically managing the relative waiting times between regular and superior services to maximize profit remains a robust finding, unaffected by the market coverage assumption. Additionally, as in the baseline model, the for-profit hospital also hires fewer superior doctors than the first-best level ($\mu_s^{PM} < \mu_s^{WM}$). This directly results from the intentionally suppressed demand for superior service.

PROPOSITION 7. The equilibrium waiting time of both regular and superior doctor services are longer in the profit-maximizing scenario than in the welfare-maximizing scenario. That is, $W_j^{PM} > W_j^{WM}$, $j = r, s$.

The results in Proposition 7 are consistent with the results in Proposition 3 in the baseline model that, the profit-driven decisions of a for-profit hospital would lead to longer waiting time and higher per capita waiting cost (i.e., $c_w W_j^{PM} > c_w W_j^{WM}$, $j = r, s$) and thus reduced efficiency for each type of service compared to the first-best level. Propositions 6 and 7 confirm the robustness of our core findings even after relaxing the full-market coverage assumption and explain why

for-profit hospitals are usually outperformed by nonprofit hospitals in accessibility, efficiency, and average service quality even with flexible market participation.

4.2 Flexible Regular Capacity

In this section, we relax the assumption of fixed regular doctor capacity and endogenize the regular service price p_r , allowing the hospital to flexibly adjust both regular doctor capacity and price as well. It is worth noting that the full-market coverage assumption requires the capacity of regular doctors to be fixed. Otherwise, the hospital would simply stop hiring regular doctors to make regular service infinitely bad and push patients to pay for the superior service. Therefore, to relax the assumption for fixed capacity on regular doctors, we must also relax the assumption of full-market coverage. As in Section 4.1, we allow flexible market participation by patients in this section. More details of the model are provided in the online appendix.

In the model studied in Section 4.2, with flexible market coverage and flexible regular capacity, the demand for regular and superior services is no longer mutually correlated. The flexibility in capacity and pricing for regular services introduces an extra degree of freedom. In what follows, we explore the robustness of the comparison results between the profit-maximizing and welfare-maximizing scenarios in the setting of flexible market participation, flexible regular capacity, and endogenized regular service price.

PROPOSITION 8. Let $(k_r^{PM}, k_s^{PM}, \mu_s^{PM})$ denote the equilibrium complexity thresholds of patient partition and the optimal capacity of superior doctor service in the profit-maximizing scenario, and let $(k_r^{WM}, k_s^{WM}, \mu_s^{WM})$ denote the equilibrium complexity thresholds of patient partition and the optimal capacity of superior doctor service in the welfare-maximizing scenario. Then we have that

- (i) fewer patients choose to get treatment in the profit-maximizing scenario, that is, $k_r^{PM} > k_r^{WM}$;
- (ii) the profit-maximizing scenario directs fewer patients to the superior doctor service and fewer patients to the regular doctor service than the socially optimal number of patients, that is, $k_s^{PM} > k_s^{WM}$ and $k_s^{PM} - k_r^{PM} < k_s^{WM} - k_r^{WM}$;
- (iii) the profit-maximizing scenario has fewer superior doctors and regular doctors than the socially optimal capacity, or $\mu_s^{PM} < \mu_s^{WM}$ and $\mu_r^{PM} < \mu_r^{WM}$.

PROPOSITION 9. The equilibrium waiting time of both regular and superior doctor services are longer in the profit-maximizing scenario than in the welfare-maximizing scenario. That is, $W_j^{PM} > W_j^{WM}$, $j = r, s$.

Consistent with result (i) in Proposition 6, the first result in Proposition 8 concludes that fewer patients would choose to get treated in the profit-maximizing scenario, indicating that compared with the first-best level, the level of service accessibility is lower in for-profit hospitals even with flexible regular

service price and capacity. Consistent with Propositions 2 and 3, results (ii) and (iii) in Propositions 8 conclude that a for-profit hospital directs fewer patients to the superior service and would hire fewer superior doctors than the first-best level, and thus confirm the robustness of our core findings as in the baseline mode. Furthermore, a for-profit hospital has more severe congestion and longer waiting times in both types of services and thus reduced efficiency for each type of service compared to the first-best level.

It is worth noting that, similar to the results in Proposition 6, due to a decrease in the total number of patients receiving treatment ($k_r^{PM} > k_r^{WM}$) and a decrease in the number of patients receiving premium services ($k_s^{PM} > k_s^{WM}$), the average healthcare quality for the entire population has decreased compared to the first-best socially optimal outcomes. In summary, Propositions 8 and 9 confirm the robustness of our core findings that for-profit hospitals underperform nonprofit hospitals in terms of average service quality, average waiting times, and per capita waiting cost, even with the most general model settings: flexible market participation accompanied by endogenized pricing and capacity decisions for both regular and premium services.

4.3 Heterogeneous Patients' Price Sensitivity

In reality, patients can be heterogeneous in multiple dimensions. In this section, we explore another dimension of patient heterogeneity: price sensitivity. Building upon the baseline model that accounts for variations in patients' healthcare needs, we introduce an additional independent layer of heterogeneity: patients' varying degrees of price sensitivity. This augmentation enriches our model and offers a more comprehensive representation of real-world complexities.

We assume that there are two types of patients, group L and group H , with price sensitivity ρ_L and ρ_H ($\rho_H > \rho_L$), respectively. The two groups are equal in size. For ease of exposition and without loss of generality, let $\rho_L = 1$ and $\rho_H = \rho > 1$. Given the for-profit hospital's decisions Δp and W_s , we denote the equilibrium patient partition thresholds of group L and H as k_L^* and k_H^* , respectively. Similar to the baseline model, k_L^* is such that $k_L^* - c_w W_r = \epsilon k_L^* - \Delta p - c_w W_s$ and k_H^* is such that $k_H^* - c_w W_r = \epsilon k_H^* - \rho \Delta p - c_w W_s$.

LEMMA 7. *For a profit-driven hospital, given any patient partition between regular and superior doctors, that is, $\frac{\lambda}{2}(k_L^* + k_H^*) = \lambda k^*$, the elasticity of patient partition to price increases with ρ .*

Lemma 7 aligns with our anticipation that when a group of patients displays higher price sensitivity, the overall elasticity of the total demand partition in response to price changes increases. This heightened elasticity makes price adjustments a more influential mechanism for managing the relative waiting times of the two service types. Consequently, we have the following result on the hospital's optimal patient partition decisions.

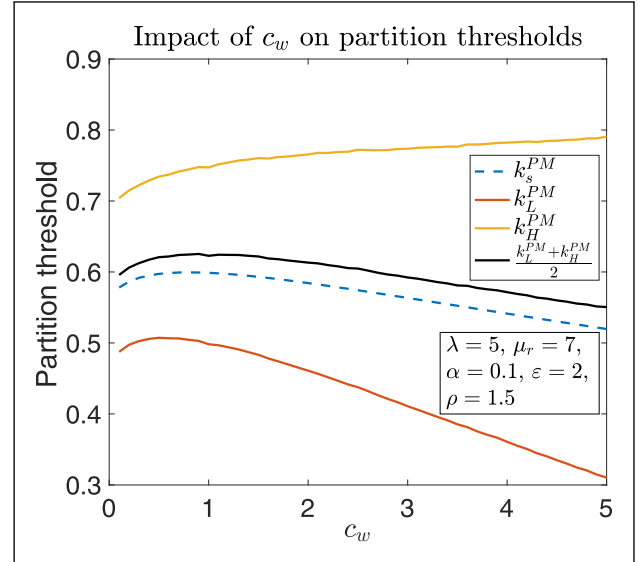


Figure 4. Partition thresholds in two groups with different price sensitivities (dashed line: baseline model).

PROPOSITION 10. *Denote k_L^{PM} and k_H^{PM} as the equilibrium patient partition thresholds given the profit-driven hospital's optimal decisions. Given that $\rho > 1$, we have $\frac{k_L^{PM} + k_H^{PM}}{2} > k^{PM}$.*

Proposition 10 implies that when certain patients exhibit high price sensitivity ($\rho > 1$), in contrast to the baseline setting where all patients have a sensitivity of $\rho_L = 1$, the for-profit hospital directs even fewer patients to the superior service than the first-best level, deviating further from the first-best results. This finding resonates with our core conclusions, emphasizing that a for-profit hospital strategically manipulates the relative waiting times between the two services to capitalize on the increased willingness of patients to pay for the superior service. As patients' price sensitivity increases, the hospital must intensify this manipulation by directing a greater number of patients towards the regular service to achieve a high premium for the superior service.

In Figure 4, we dissect the aggregate demand for regular services based on various subgroups of price sensitivity under the for-profit hospital's optimal pricing and capacity decisions. We observe that the overall demand for regular services ($\frac{k_L^{PM} + k_H^{PM}}{2}$) follows a similar trend with respect to patients' time sensitivity c_w , akin to the baseline model. Additionally, we note that the for-profit hospital tends to allocate a smaller proportion of price-sensitive patients to the superior service, while directing a majority of price-insensitive patients towards the same when patients' time sensitivity is high. This observation implies that in scenarios where patients exhibit multiple dimensions of heterogeneity, the trade-offs between price and waiting time vary across different patient groups. The for-profit hospital adeptly leverages these distinctions to make optimal decisions.

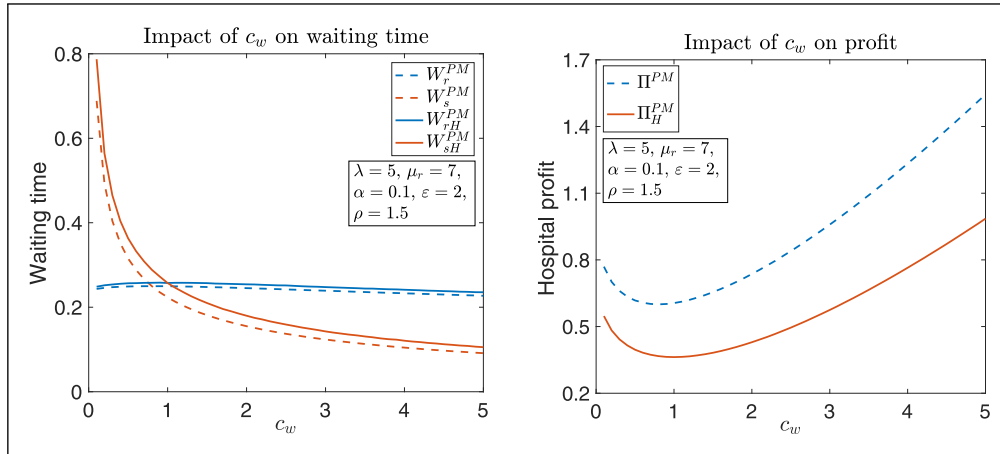


Figure 5. Effect of time value on waiting time and profit (dashed lines: baseline model).

In the extension model featuring patient heterogeneity in price sensitivity, let Δp_H^{PM} denote the for-profit hospital's optimal decision on superior doctor price premium and let W_{sH}^{PM} denote the waiting time for superior service in equilibrium. For the sake of conciseness in the main body, we first establish the robustness of our core findings in the baseline model through numerical illustrations (Figure 5) and verify that with consumer heterogeneity in price sensitivity, the hospital's profit first decreases with c_w when $W_{rH}^{PM} \leq W_{sH}^{PM}$, and then increases with c_w when $W_{rH}^{PM} > W_{sH}^{PM}$ (detailed proofs are provided in online appendix).

PROPOSITION 11. *When some patients are highly price sensitive, the equilibrium waiting time of both services are longer compared to the homogenous price sensitivity case, that is, $W_{jH}^{PM} > W_j^{PM}$, $j = r, s$. Additionally, the equilibrium price premium is smaller; that is, $\Delta p_H^{PM} < \Delta p^{PM}$.*

Given that the settings regarding the distribution of patients' healthcare needs and their waiting costs remain consistent with the baseline model, the welfare-maximizing outcomes will mirror those in the baseline model. Therefore, Proposition 11 indicates that when some patients exhibit higher price sensitivity than the baseline level, the profit-driven distortion in waiting time ($W_{jH}^{PM} > W_j^{PM} > W_j^{WM}$, $j = r, s$) and per capita waiting cost for patients intensify. This is because patients' increased price sensitivity reduces their willingness to pay, compelling the hospital to reduce the number of superior doctors in their staffing. As a result, the efficiency of the healthcare system is even lower.

5 Conclusion

In this study, we analyze the pricing and capacity decisions for two-tier healthcare systems that offer both basic and premium services, primarily through two types of doctor services:

superior doctor services and regular doctor services. Waiting time is endogenously determined by patient partition in equilibrium. By comparing the optimal decisions in both the profit-maximizing scenario and the welfare-maximizing scenario, we analyze the welfare loss caused by profit-driven healthcare facilities and decompose the sources of such welfare loss. Additionally, we demonstrate the robustness of our results by extending the baseline model to a series of more realistic and practical extensions.

We find that compared with welfare-driven hospitals, profit-driven hospitals invest less in superior doctor capacity in both full market coverage and flexible market participation scenarios. At equilibrium, fewer patients are treated by the superior doctor service in profit-driven hospitals, resulting in a reduced average quality of service. This is primarily because the profit-driven hospitals are incentivized to differentiate between the waiting times of the two service types to induce patients to opt for the superior doctor service and pay a higher price. Intuitively, service value has two dimensions: time and service quality, and price is used as a tool for managing the waiting time in the model when waiting times are endogenously determined. When the profit-driven hospital increases the premium for the superior doctor service, some patients switch to the regular doctor service which makes the superior doctor service queue shorter. At the same time, as the superior service queue becomes shorter, the patients who choose to stay with the superior doctor service are willing to pay an even higher premium for this service. This interplay between the roles of price as a waiting time management tool and a profit channel leads to the inefficiency distortions in the profit-driven hospital's decisions.

Our results also show that if the inefficiency in patient partition is corrected, the profit-driven hospital would choose the welfare-maximizing level of superior doctor capacity and thus achieve socially optimal results. However, we also find that common types of policy regulations, such as price ceilings and capacity regulation, cannot fully correct the

inefficiencies in patient partition and capacity decisions of the profit-driven hospital.¹ When price is capped at the socially optimal level, the profit-driven hospital underinvests in superior doctor capacity, which leads to fewer patients choosing the superior service than in the welfare-maximizing scenario. When superior doctor service capacity is required to be no lower than the socially optimal level, the hospital sets a higher premium for the superior doctor service than in the welfare-maximizing scenario, which causes more than the socially optimal level of patients to opt for the regular doctor service.

Our results fill a gap in the literature by explaining the findings summarized in Rosenau and Linder (2003) and Kruse et al. (2018) that for-profit hospitals often underperform non-profit hospitals in healthcare service accessibility, average quality of service, and efficiency from a modeling perspective. Additionally, our results provide important insights for healthcare facilities managers and policymakers. Profit-driven hospital should understand that (1) price can be used as a tool for managing the waiting time and affects the trade-off between service quality and waiting time for patients, and (2) a higher value of waiting time does not always lead to higher profit. Policymakers aiming to improve social welfare should focus on correcting the inefficiency in the patient partition threshold which determines the match between patient illness complexity and doctor types.

This research can be extended in several ways. First, for model tractability, we assume patient illness complexity follows a uniform distribution and the mean service time of doctors is independent of patient illness complexity. In reality, for some diseases, a majority of the patients may have low complexity while a smaller percentage of the patients suffer the illness of high complexity. The service time may also positively relate to illness complexity. Therefore, this research can be extended by relaxing these assumptions and considering other patient illness complexity distributions or allowing the dependence of service time on illness complexity. Second, the hospital we consider offers two types of doctor services (regular and superior) with a difference in service quality denoted by ε . In practice, the heterogeneity in service quality may be more complicated. For example, even within the superior type of doctors, depending on individual factors such as training and experience, the service quality from each doctor can be quite different. Therefore, modeling heterogeneous service quality in a more comprehensive way should be a meaningful direction to explore. Lastly, in our model, the hospital determines the price of each type of service. An interesting direction to look at for future research would be allowing superior doctors of different quality to price the service on their own.

Appendix A. List of Notations

Table A1. Model notations.

λ	mean arrival rate of patients
k	complexity of a patient's condition, $k \sim U(0, 1)$
V	base value if a patient is treated
$R(k)$	treatment value function of a patient with complexity k
ε	service quality factor of superior doctors
c_w	opportunity cost per unit of time of a patient
α	unit cost of superior doctor capacity
ρ	the price sensitivity of highly price-sensitive patients, $\rho > 1$
W_r, W_s	expected waiting time of each service at regular and superior doctors, respectively
p_r, p_s	price for regular and superior doctors, respectively
$\Delta p = p_s - p_r$	the difference between the prices of superior and regular doctor service
μ_r, μ_s	mean service rates of regular and superior doctors, respectively
WC_j^{PM}	total waiting cost of all patients who choose the type j service in the profit-maximizing scenario, $j = r, s$
WC_j^{WM}	total waiting cost of all patients who choose the type j service in the welfare-maximizing scenario, $j = r, s$

Acknowledgments

We would like to express our gratitude for the helpful feedback provided by the department editor, senior editor, and two anonymous reviewers.


Declaration of Conflicting Interests


The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported in part by the National Natural Science Foundation of China 72101092, the Guangdong Basic and Applied Basic Research Foundation 2020A1515110239.

ORCID iDs

Shengya Hua  <https://orcid.org/0000-0002-8256-7873>

Xin Zhai  <https://orcid.org/0000-0002-0401-6670>

Supplemental Material

Supplemental material for this article is available online (doi: 10.1177/10591478241238969).

Note

1. Numerical studies on how these interventions enhance social welfare are provided in the online appendix.

References

- Anderson SP and Renault R (2009) Comparative advertising: Disclosing horizontal match information. *The RAND Journal of Economics* 40(3): 558–581.
- Armstrong M and Wright J (2007) Two-sided markets, competitive bottlenecks and exclusive contracts. *Economic Theory* 32(2): 353–380.
- Australian Capital Territory (ACT) (2023) What are nurse/midwife to patient ratios? Available at URL <https://www.health.act.gov.au/health-professionals/nursing-and-midwifery-office/nurse/midwife-patient-ratios> (accessed date February 1, 2024).
- Barber SL, Lorenzoni L and Ong P (2019) Price setting and price regulation in health care: Lessons for advancing Universal Health Coverage, OECD-WHO.
- Bimpikis K and Markakis MG (2019) Learning and hierarchies in service systems. *Management Science* 65(3): 1268–1285.
- Church J and Gandal N (1992) Network effects, software provision, and standardization. *The Journal of Industrial Economics* 40(1): 85–103.
- Coban E, Heching A and Scheller-Wolf A (2019) Service center staffing with cross-trained agents and heterogeneous customers. *Production and Operations Management* 28(4): 788–809.
- Dai T and Tayur S (2020) OM Forum—healthcare operations management: A snapshot of emerging research. *Manufacturing & Service Operations Management* 22(5): 869–887.
- Dewan S and Mendelson H (1990) User delay costs and internal pricing for a service facility. *Management Science* 36(12): 1502–1517.
- Du AY, Das S, Gopal RD, et al. (2014) Optimal management of digital content on tiered infrastructure platforms. *Information Systems Research* 25(4): 730–746.
- Guo P and Zhang ZG (2013) Strategic queueing behavior and its impact on system performance in service systems with the congestion-based staffing policy. *Manufacturing & Service Operations Management* 15(1): 118–131.
- Hasija S, Pinker EJ and Shumsky RA (2005) Staffing and routing in a two-tier call center. *International Journal of Operational Research* 1(1/2): 8–29.
- Hassin R (2016) *Rational Queueing*. London: CRC Press.
- Hassin R and Haviv M (2003) *To queue or not to queue: Equilibrium behavior in queueing systems*. Boston, MA: Kluwer Academic Publishers, Kluwer.
- Jeon DS, Laffont JJ and Tirole J (2004) On the “receiver-pays” principle. *RAND Journal of Economics* 35(1): 85–110.
- Kruse FM, Stadhouders NW, Adang EM, et al. (2018) Do private hospitals outperform public hospitals regarding efficiency, accessibility, and quality of care in the European Union? A literature review. *The International Journal of Health Planning and Management* 33(2): e434–e453.
- Lee HH, Pinker EJ and Shumsky RA (2012) Outsourcing a two-level service process. *Management Science* 58(8): 1569–1584.
- Maglaras C, Yao J and Zeevi A (2017) Optimal price and delay differentiation in large-scale queueing systems. *Management Science* 64(5): 2427–2444.
- Marchand MG (1974) Priority pricing. *Management Science* 20(7): 1131–1140.
- National Informatics Centre (NIC) (2019) About the Department. Available at URL <http://pharmaceuticals.gov.in/about-department> (accessed date February 1, 2024).
- Organization for Economic Co-operation and Development (OECD) (2022) Doctors (indicator). Available at URL <https://data.oecd.org/healthres/doctors.htm> (accessed date July 26, 2022).
- Propper C (1990) Contingent valuation of time spent on NHS waiting lists. *Economic Journal* 100(400): 193–199.
- Qian Q, Guo P and Lindsey R (2017) Comparison of subsidy schemes for reducing waiting times in healthcare systems. *Production and Operations Management* 26(11): 2033–2049.
- Rajan B, Tezcan T and Seidmann A (2019) Service systems with heterogeneous customers: Investigating the effect of telemedicine on chronic care. *Management Science* 65(3): 1236–1267.
- Roberts A (2023) Nurse-Patient Ratios: These States Have These Controversial Policies in Place. Available at URL <https://nursejournal.org/articles/nurse-patient-ratios/> (accessed date February 1, 2024).
- Rosenau PV and Linder SH (2003) Two decades of research comparing for-profit and nonprofit health provider performance in the United States. *Social Science Quarterly* 84(2): 219–241.
- Saghafian S, Hopp WJ, Iravani SMR, et al. (2018) Workload management in telemedical physician triage and other knowledge-based service systems. *Management Science* 64(11): 5180–5197.
- Shumsky RA and Pinker EJ (2003) Gatekeepers and referrals in services. *Management Science* 49(7): 839–856.
- Stidham S (1992) Pricing and capacity decisions for a service facility: Stability and multiple local optima. *Management Science* 38(8): 1121–1139.
- Stidham S (2009) *Optimal Design of Queueing Systems*. London: CRC Press.
- Tarakci H, Ozdemir Z and Sharafali M (2009) On the staffing policy and technology investment in a specialty hospital offering telemedicine. *Decision Support Systems* 46(2): 468–480.
- World Health Organization (WHO) (2022) Universal Health Coverage. Available at URL <https://www.who.int/health-topics/universal-health-coverage/> (accessed date July 26, 2022).
- Zhang Z and Yin X (2021) Designing a sustainable two-tier service system with customer’s asymmetric preference for servers. *Production and Operations Management* 30(11): 3856–3880.

How to cite this article

Hua S, Lei Y and Zhai X (2024) Pricing and Capacity Design for Profit-Driven and Welfare-Driven Healthcare Providers. *Production and Operations Management* 33(4): 1014–1030.