https://doi.org/10.1287/msom.2017.0690

The Impact of Reimbursement Policy on Social Welfare, Revisit Rate and Waiting Time in a Public Healthcare System: Fee-for-Service vs. Bundled Payment

Pengfei Guo

Faculty of Business, the Hong Kong Polytechnic University, Hong Kong, pengfei.guo@polyu.edu.hk

Christopher S. Tang

Anderson School of Management, University of California, Los Angeles, California 90095, chris.tang@anderson.ucla.edu

Yulan Wang

Faculty of Business, the Hong Kong Polytechnic University, Hong Kong, yulan.wang@polyu.edu.hk

Ming Zhao

School of Economics and Management, University of Electronic Science and Technology of China, Chengdu, China 610054; and

Faculty of Business, the Hong Kong Polytechnic University, Hong Kong, lighting.zhao@connect.polyu.hk

Abstract

This paper examines the impact of two reimbursement schemes, Fee-for-Service and Bundled Payment, on the social welfare, the patient revisit rate, and the patient waiting time in a public healthcare system. The two schemes differ on the payment mechanism: under the fee-for-service scheme, the healthcare provider receives the payment each time a patient visits (or revisits) whereas, under the bundled payment scheme, the healthcare provider receives a lump sum payment for the entire episode of care regardless of how many revisits a patient incurs. By considering the quality-speed tradeoff (i.e., a higher service speed reduces service quality, resulting in a higher revisit rate), we examine a three-stage Stackelberg game to determine the patients' initial visit rate, the service provider's service rate (which affects the revisit rate), and the funder's reimbursement rate. This analysis enables us to compare the equilibrium outcomes (social welfare, revisit rate and waiting time) associated with the two payment schemes. We find that when the patient pool size is large, the bundled payment scheme dominates the fee-for-service scheme in terms of higher social welfare and a lower revisit rate, whereas the fee-for-service scheme prevails in terms of shorter waiting time. When the patient pool is small, the bundle payment scheme dominates the fee-for-service scheme in all three performance measures.

Keywords: Healthcare operations, Fee-for-Service, Bundled Payment, Queueing.

1 Introduction

Public healthcare systems are facing many challenges, such as escalating operating costs, deteriorating service quality and lengthening waiting times. For example, the cost of public healthcare insurance for the average Canadian family increased by 48.5% from 2005 to 2015, which is 1.6 times the national salary increase over the same period (Palacios et al. 2015). At the same time, Canadian patients often wait for 18.2 weeks for elective treatments (Barua and Fathers 2014). In the United Kingdom (UK), waiting times for elective surgery are considered by the public to be the second most important failing of the public healthcare system: the average waiting time is 95 days for knee replacements, 68.8 days for cataract surgery, and 80.7 days for hernia repairs (Campbell 2014). In Hong Kong, the waiting time for cataract surgery is longer than eight months (Hospital Authority 2016). Because excessive long waiting times cause patient dissatisfaction, some public healthcare systems (such as in the UK) include waiting time (especially for elective surgeries) as a key performance measure (Dimakou 2013), and others (such as in South Australia) are committed to reducing waiting time (Official Website of the South Australian Government).

Clearly, increasing the (public healthcare) system capacity can reduce waiting times and improve patient satisfaction. However, on one hand, expanding the capacity of a healthcare system is very costly and time-consuming (Guo et al. 2016); on the other hand, the public healthcare systems in many countries face limited budgets and large patient volume. Therefore, given a fixed capacity, short- to medium-term solutions are of great interest to the policy makers in healthcare systems. In this work, we consider such a mechanism, a payment scheme. Our motivation is based on the fact that many healthcare professionals believe that an effective reimbursement scheme can induce healthcare providers (HCPs) to reduce waiting time, contain costs and improve service quality. Currently, the predominant scheme is called Fee-for-Service (FFS) under which an HCP receives payment each time a patient is admitted (or re-admitted). The FFS scheme creates incentives for HCPs to urge their doctors to rush through their appointments so as to treat more patients per day (Rabin 2014), even though it is known to be an effective scheme for reducing waiting time (Blomqvist and Busby 2013). Without resolving patients' problems completely, more revisits will ensue (van der Linden et al. 2011). In addition, the FFS scheme creates major concerns, including: (1) excessive treatments (Davis 2007); (2) more revisits (Fenter and Lewis 2008); and (3) low service quality at a high cost (Calsyn and Lee 2012).

To improve service quality and contain costs, the Centers for Medicare and Medicaid Services (CMS) in the United States is gradually shifting from the FFS scheme to the Bundled Payment (BP) scheme under which the HCP receives a lump sum payment for the entire episode of care (within a specified time window), regardless of the number of times a patient

is readmitted (Tsai et al. 2015). A recent study found that, relative to the FFS scheme, the BP scheme can reduce the cost per episode of care by 3% (Japsen 2015). At the same time, Ontario (Canada) has been examining the effectiveness of the BP scheme since 2011 (Ministry of Health and Long-Term Care 2015), and the Australian government was considering the BP scheme in 2015 (Australian Primary Health Care Advisory Group Report 2015).

While the BP scheme has received considerable attention, many public healthcare systems continue to operate under the FFS scheme because the underlying implications are not well understood. Therefore, it is important to characterize the effects of these two schemes on certain performance measures, including social welfare and service quality (revisits and waiting time). In this paper, we compare these performance measures associated with the FFS and BP schemes in relation to providing outpatient elective care services in a public healthcare system that consists of a funder, an HCP and a population of patients. To facilitate our comparative analysis, we use a three-stage Stackelberg game to capture the dynamic interactions among all three parties.

The funder, who acts as the first leader, determines the reimbursement rate to maximize the social welfare. The social welfare here is defined from a wider scope: the funder cares about not only the welfare of those patients who seek treatments from the public healthcare system but also the burden on outside systems caused by the overflow of those patients who have to seek treatments elsewhere. Given the reimbursement rate, the HCP provides elective care service and acts as the second leader who decides on the service rate to maximize its profit, where a higher service rate yields a higher revisit rate. Finally, given the HCP's service rate, each patient decides whether or not to seek elective care from the HCP by taking other patients' visits into consideration. Therefore, the patient's visit rate is *endogenously* determined according to a Nash equilibrium. Actually, for elective care service in a public system, each patient can seek help from the HCP or elsewhere. For example, starting in 2013 and partly in order to address the issue of long waiting time, the European Union decided to grant European citizens the freedom to choose the member-state from which they receive care while being entitled to reimbursement from their home insurance systems (Andritsos and Tang 2014).

Embedded in our three-stage Stackelberg game is an M/M/1 queueing model with endogenous patient arrivals and revisits. This queueing model enables us to determine the service rate (decided by the HCP) and the corresponding patient arrival rate, revisit rate, waiting time and social welfare for any given reimbursement rate (specified by the funder) under both FFS and BP schemes. Our analysis captures three issues. First, the patients' joining (or balking) decisions are rational; they take other patients' joining decisions into consideration when deciding whether or not to join. When the patients are homogeneous,

we determine a mixed-strategy Nash equilibrium that is characterized by the probability of "joining" for each patient (Hassin and Haviv 2003). By considering this joining probability, we can determine the endogenous patient arrival rate in equilibrium that depends on the HCP's service rate. Second, because the endogenous patient arrival rate depends on the waiting times, the HCP's service rate decision process is complex; a higher service rate does not necessarily result in shorter waiting times due to a higher revisit rate. Third, the size of the patient pool plays an important role in explaining the implications of the FFS and BP schemes. Specifically, when the size of the pool is large, reducing revisits (by reducing service rate) may not reduce waiting times because more patients will join the system. However, when the pool is small so that the HCP can serve all patients in the pool, reducing revisits can reduce waiting times.

After solving the HCP's optimal service rate decision, we can then consider the funder's decision on the subsidy scheme with an objective of maximizing the social welfare. We find that in a large-sized market where balking patients exist, maximizing the social welfare is degenerated to maximizing the initial visit rate, that is, increasing the service accessibility. We find that under both FFS and BP reimbursement schemes, a higher reimbursement rate will lead to a lower service rate, which will result in higher quality but also a higher level of congestion. That is, the quality and congestion level cannot be simultaneously improved by increasing the subsidy amount. However, the two criteria can be simultaneously improved in a small-sized market where all patients are served.

By comparing the equilibrium outcomes (the social welfare, the revisit rate and waiting time) associated with these two schemes, we find that the dominance of one scheme over the other depends heavily on the size of the patient population, as follows:

1. When the patient population is sufficiently large, the BP scheme dominates the FFS scheme in terms of higher social welfare and a lower revisit rate. However, the FFS scheme outperforms the BP scheme in terms of both shorter waiting time per visit and shorter total waiting time in the system.

2. When the patient population is sufficiently small, the BP scheme dominates the FFS scheme in terms of higher patient utility, a lower revisit rate and shorter waiting times.

3. When the patient population is medium, we identify exact conditions under which the BP scheme and the FFS scheme yield identical performance.

This paper makes two contributions to the healthcare operations literature. First, our paper represents a new attempt to examine the implications of two reimbursement schemes by incorporating issues of endogenous patient elected visits and random revisits arising from

a public healthcare system that provides elective care. Second, our analysis provides insights regarding the conditions under which one scheme outperforms the other in terms of social welfare, the revisit rate and waiting time. Our finding has the following implications for the government (or the CMS): the BP scheme dominates the FFS scheme in a system with less congestion. However, in a congested system, the government should be aware of the following trade-off: the BP scheme is more effective in reducing the revisit rate, but the FFS scheme is more effective in reducing the waiting time. In summary, switching from the FFS scheme to the BP scheme can create a *polarized effect*, that is, it can make a congested system even more congested and a light-traffic system even lighter.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature. In Section 3, we present our queuing model and establish some preliminary results. In Section 4, we analyze our three-stage Stackelberg game by determining the equilibrium outcomes associated with the FFS and BP schemes when the patient population is large, while in Section 5, we compare the equilibrium outcomes under two schemes when the patient population is small. We enrich our model in Section 6. Concluding remarks are provided in Section 7. Proofs (except that of Proposition 7) are relegated to the online Appendix. Some analysis on cases not presented here can be found in an early version of this paper, Guo et al. (2017).

2 Literature Review

As public funders in different countries are contemplating whether they should change the payment scheme from FFS to BP, researchers are developing different models to compare the performance measures associated with these two schemes. For example, Adida et al. (2017) consider a healthcare system in which a risk-averse HCP can select the type of patients to admit and decide the treatment intensity for each admitted patient. By analyzing a two-stage model, the authors examine the impact of the FFS and BP schemes on patient selection and treatment intensity. They find that the HCP has the incentive to provide excessive treatments under FFS and to incur suboptimal patient selection under BP. To alleviate the shortcomings of FFS and BP, the authors propose two alternative payment systems that may induce system optimal decisions. In a different setting, Andritsos and Tang (2015) consider a situation in which the patient care can be *co-managed* by the HCP and the patient so that the revisit depends on the effort exerted by both the HCP and the patient. They show that the BP scheme outperforms the FFS scheme in terms of patient welfare because the BP scheme can induce the HCP and the patient to exert more revisit-reduction efforts. Gupta and Mehrotra (2015) examine the BP scheme for Care Improvement (BPCI) initiative

initiated by the CMS. The BPCI invites HCPs to propose bundles of services along with target payments per episode, quality targets, etc. By considering the proposed selection process adopted by the BPCI, they derive an optimal strategy for the CMS to consider.

Although we also focus on the comparison of performance measures associated with the FFS and BP schemes, our paper complements the above work in the following manner. First, unlike the setting examined in Adida et al. (2017) in which the HCP selects which type of patients to admit, we consider a situation in which patients are sensitive to waiting time and can select not to seek elective care from the public HCP so that the patients' arrival rate is endogenously determined by the patients (not the HCP). Correspondingly, Adida et al. (2017) and our paper use different measurements to evaluate the performance of FFS and BP schemes; Adida et al. (2017) focus on treatment intensity and patient selection, but we focus on the revisit rate, waiting time, and social welfare. Second, unlike those two-stage models developed by Adida et al. (2017) and Andritsos and Tang (2015) in which a patient can only be re-admitted at most once, we use a queueing model with random revisits over time to determine the patient's total waiting time in the system. Third, while Gupta and Mehrotra (2015) examine the auction-like mechanism adopted by the CMS, we are interested in comparing the social welfare, the revisit rate, and waiting times associated with the FFS and BP schemes.

Our paper is also related to the healthcare operations management literature that examines the performance of different payment schemes. Specifically, there is a stream of literature that examines various performance-based payment schemes. So and Tang (2000) examine the impact of an outcome-oriented drug reimbursement policy on the patients' health. By using a dynamic principal-agent model, Fuloria and Zenios (2001) find that a patient outcome-based reimbursement scheme is effective for improving service quality. Lee and Zenios (2012) empirically show that an evidence-based payment system with risk adjustment can induce the HCP to improve its service quality. Jiang et al. (2012) study the performance-based contract for an outpatient service and show that a nonlinear performance contract can achieve the second-best performance and coordinate the service supply chain. Bavafa et al. (2013) study the patients' and physicians' responses to two care delivery innovations in primary care. Specifically, they consider e-visits and non-physician providers under both the FFS and capitation compensation schemes, where the capitation reimbursement scheme pays physicians a fixed amount for each enrolled patient during a fixed period of time, no matter whether the patient seeks care or not. Guo et al. (2016) investigate the efficiency of the conditional and unconditional subsidy schemes in healthcare systems. Dai et al. (2017) investigate physicians' imaging test ordering decisions. They show that the insurance coverage structure is the key driver of the physicians' overtesting behavior. Other

research papers in this stream include Plambeck and Zenios (2000), Yaesoubi and Roberts (2011) and Ata et al. (2013). However, none of these works considers the BP scheme.

Besides the healthcare operations management literature, our paper is related to the queueing literature that examines the issue of speed-quality trade-off (i.e., the service quality depends on the service rate) so that the service rate is endogenously determined. First, when the service quality depends on the service rate, Hopp et al. (2007) find that capacity expansion can make waiting time longer. Second, when the service quality is decreasing in the service rate and the arrival rate is endogenously determined by the customers, Anand et al. (2011) investigate the optimal pricing strategy for the service provider who provides customer-intensive services. They find that a larger number of servers leads to a higher price and a lower service speed. In the same vein, Kostami and Rajagopalan (2014) analyze the quality-speed trade-off in a dynamic setting. Tong and Rajagopalan (2014) compare the fixed fee and time-based fee schemes and identify conditions under which one scheme dominates the other. Li et al. (2016) consider the quality-speed trade-off with bounded rational customers. While the above papers examine the issue of quality-speed trade-off in a general context, there are papers that deal with this issue in industry-specific contexts, including diagnostic services (Paç and Veeraraghavan 2010, Wang et al. 2010, Alizamir et al. 2013), service quality variability (Xu et al. 2015), call center (de Vericourt and Zhou 2005, Hasija et al. 2009) and healthcare staffing (Yom-Tov and Mandelbaum 2014). While de Vericourt and Zhou (2005), Chan et al. (2014) and Yom-Tov and Mandelbaum (2014) consider returning customers, we consider the case where the arrival process is endogenously determined by the patients while the revisit rate is endogenously determined by the HCP (via its selection of service rate). Moreover, our focus is on the comparison of various performance measures associated with the FFS and BP schemes, and our results enable us to specify the conditions under which one scheme dominates the other.

3 Model Preliminaries

We now present our model that involves a pool of homogeneous patients, an HCP and a public funder, along with certain assumptions. We shall discuss the limitations of our model in the concluding section. Consider a public healthcare system consisting of a funder who sets the reimbursement rate subject to a limited budget, an HCP who determines its service rate, and a pool of homogeneous patients who decide whether or not to seek elective treatments from the public HCP. The assumption about homogeneous patients is reasonable given that in many countries, such as France, Germany and the United States, patients are classified into different diagnosis-related groups according to their respective symptoms, and

the patients in the same group demand similar resources and services (e.g., Street et al. 2011). The HCP provides a single outpatient elective treatment (e.g., hernia repairs). We model the HCP operation as an M/M/1 queue with random revisits (via Bernoulli trials). Specifically, we consider the case when *potential* patients "arrive" at the HCP according to a Poisson process with a rate of Λ. In our queueing model, we refer to patient arrivals as those patients who wish to make appointments with the HCP, and we refer to the waiting time in the queue as the waiting time for an appointment. (In other words, we do not consider the waiting time during an appointment at the clinic.) Some patients join the system while others balk. The arrival rate formed by those joined patients is called the initial arrival rate and denoted by λ.

In our queueing model, the HCP serves patients on a first-come-first-serve (FCFS) basis, where the service takes an exponentially distributed service time at rate μ. Both the Poisson arrival process and exponential service time have been well tested in the healthcare operations management literature. For instance, Kim et al. (1999) empirically verify that the arrival process at a hospital intensive care unit follows a Poisson process, and the service time follows an exponential distribution. Akin to Anand et al. (2011), we shall assume that the HCP's capacity (i.e., the number of doctors) is fixed. In practice, the capacity change due to increasing the number of doctors is costly and time consuming. For example, the supply of primary care physicians in the United States, measured by the number per 100,000 population, remained stable from 2002 to 2012 (Hing and Hsiao 2014). However, the HCP can change its service rate μ by adjusting the service time per patient so as to maximize its profit, in the same spirit as that of Andtritsos and Tang (2014). To capture the quality-speed trade-off, we shall consider the case in which the patient revisit is more likely to occur when the HCP increases its service rate. For example, for outpatient hernia repair operations, van der Linden et al. (2011) report that a shorter operating time can lead to a higher recurrence rate of re-operation (i.e., revisit).

After discharge, each patient is either cured or revisits with probability $\delta(\mu)$, where $\delta(\mu)$ represents the average statistics across all patients. For tractability, we assume that each revisited patient will join the end of the queue (i.e., the revisited patients do not have priority). This assumption is reasonable for (non-urgent) elective outpatient healthcare services such as hernia repair operations. We also make the following assumptions about the revisit rate $\delta(\mu)$:

Assumption 1: The revisit rate $\delta(\mu)$ is increasing in the service rate μ, where $\delta(\mu) \in [0, 1]$, $\delta(0) = 0$, and $\delta(\infty) = 1$.

Assumption 2: The cure rate $(1 - \delta(\mu))$ is logconcave in μ; that is, $\log(1 - \delta(\mu))$ is concave so that $g(\mu) = \delta'(\mu)/(1 - \delta(\mu))$ is increasing in μ.

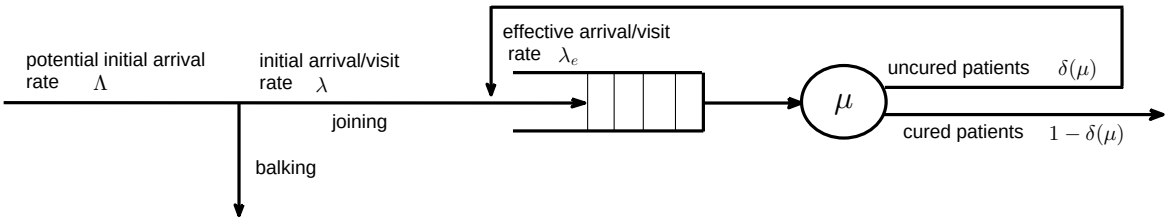

Figure 1: A schematic of the model

Assumption 1 is consistent with the empirical evidence from the outpatient practice that the revisit rate is increasing in the service rate μ for elective outpatient care such as outpatient hernia repair operations (van der Linden et al. 2011). By noting that the elasticity of the cure rate $(1-\delta(\mu))$ equals $\mu g(\mu)$, assumption 2 guarantees that the cure rate is more sensitive to the change in the service rate when the service rate is larger. Observe that the logistic function $\delta(\mu) = 1/(1+e^{-a\mu+b})$ with parameters $a > 0$ and $b > 0$ satisfies both assumptions, where the logistic function is a standard approach to measure the relationship between the revisit rate and other variables in the healthcare economics literature (e.g., Fethke et al. 1986, Morrow-Howell and Proctor 1993).

Based on the above assumptions, we can model the healthcare system as an M/M/1 queue with random revisits (via Bernoulli trials), as depicted in Figure 1. Note that the service rate for new patients and that for revisit patients are assumed to be the same. This assumption is reasonable for outpatient elective care services (such as hernia repair operations) where the appointment block for each patient is normally fixed, regardless of whether the patient is new or revisiting. For example, in the UK, physicians allocate almost the same amount of time for the initial visits and the follow-up ones (Konrad et al. 2010).

3.1 Cure Service Rate

Observe from Figure 1 that the probability that a patient is cured after a visit is equal to $1-\delta(\mu)$, where μ is the HCP's service rate. Therefore, $\mu(1-\delta(\mu))$ is the effective service rate that the HCP cures its patients. This observation motivates us to introduce a term that we refer to as the *cure service rate* $o(\mu)$, where

$$o(\mu) = \mu \cdot (1-\delta(\mu)). \tag{1}$$

As we shall see in our subsequent analysis, the cure service rate $o(\mu)$ allows us to interpret our results intuitively.

Let μ^o be the service rate that maximizes the cure service rate $o(\mu)$; that is, $\mu^o =$

$argmax\{o(\mu) : \mu \geq 0\}$. Using the first-order condition along with assumption 2, we get the following result.

Lemma 1. *The cure service rate $o(\mu)$ is unimodal in μ. The optimal μ^o is attained when the elasticity of the cure rate equals 1; that is, when*

$$\mu^o \cdot g(\mu^o) = 1. \tag{2}$$

In addition, the cure service rate $o(\mu)$ is concave in μ for $\mu \leq \mu^o$.

Lemma 1 shows that the cure service rate $o(\mu)$ has a unique mode μ^o that has the "elasticity" of cure rate $1 - \delta(\mu)$ (i.e., $\mu \cdot g(\mu)$ equals one). This result can be explained by using the following intuition. When $\mu \cdot g(\mu) < 1$, that is, when the cure rate $1 - \delta(\mu)$ is inelastic, $1 - \delta(\mu)$ changes slowly so that an increase in the service rate μ will cause a net increase in the cure service rate $o(\mu)$. By using the same logic, an increase in the service rate μ will cause a net decrease in the cure service rate $o(\mu)$ when $\mu \cdot g(\mu) > 1$. Consequently, the optimal point is attained at the service rate that has $\mu \cdot g(\mu) = 1$.

3.2 Total Waiting Time

By considering the queueing network as depicted in Figure 1, we now determine the *total waiting time* that a patient spends in the system before being cured. Here, the total waiting time includes the waiting time of the initial visit and the waiting time of all potential subsequent revisits during a *medical episode*. Let λ and λ_e denote the patients' initial arrival rate (i.e., the arrival rate of newly admitted patients) and the patients' effective arrival rate (that includes initial visits and all subsequent revisits), respectively. To be consistent with the terminology used in the healthcare industry, we shall refer the effective arrival rate of newly admitted patients as the "initial visit rate" throughout this paper. In steady state, the departure rate of the system is equal to the effective arrival rate λ_e, which, in turn, equals the sum of the initial arrival rate λ and the arrival rate associated with the revisits (which is equal to $\delta(\mu) \cdot \lambda_e$). Therefore,

$$\lambda_e = \lambda + \delta(\mu) \cdot \lambda_e \quad \Rightarrow \quad \lambda_e = \frac{\lambda}{1 - \delta(\mu)}. \tag{3}$$

Given μ, let N represent the number of visits that a patient endures before being cured. It can be shown that the expected number of visits that a patient endures before being cured, denoted by $n(\mu)$, can be expressed as (see Ross 2007, Example 2.18)

$$n(\mu) = E[N] = \frac{1}{1 - \delta(\mu)}. \tag{4}$$

From (3) and (4), we have $\lambda_e = n(\mu) \cdot \lambda$, which implies that the effective arrival rate equals the initial arrival rate λ times the expected number of visits per medical episode $n(\mu)$.

By considering an M/M/1 queue with instantaneous Bernoulli feedback (Ross 2007), it is well known that the average number of customers in the system is equal to $L = \frac{\lambda_e}{\mu - \lambda_e} = \frac{\lambda}{o(\mu) - \lambda}$. Note that our analytical results continue to hold when we relax the assumption of instantaneous feedback as long as uncured patients stay at home for an exponentially distributed amount of time before revisiting the system. Let W and T denote the expected waiting time per visit and the expected total waiting time per medical episode, respectively. Using Little's law, we have $L = \lambda_e \cdot W = \lambda \cdot T$. By combining these two observations, we get:

$$W(\lambda, \mu) = \frac{1}{\mu - \lambda_e} = \frac{1 - \delta(\mu)}{o(\mu) - \lambda}, \tag{5}$$

$$T(\lambda, \mu) = \frac{1}{o(\mu) - \lambda}. \tag{6}$$

Note that we can interpret T given in (6) as the expected waiting time associated with the classic M/M/1 queue with a corresponding arrival rate λ and service rate $o(\mu)$.

3.3 Patient Utility

For any given service rate μ, a waiting time-sensitive patient who seeks the treatment from the HCP derives her utility $U(\lambda, \mu)$ as follows:

$$U(\lambda, \mu) = R - [n(\mu) \cdot t + \theta \cdot T(\lambda, \mu)] = R - \frac{t}{1 - \delta(\mu)} - \frac{\theta}{o(\mu) - \lambda}, \tag{7}$$

where $n(\mu)$ is the number of visits a patient expects to experience per medical episode given in (4), $T(\lambda, \mu)$ is the expected total waiting time per medical episode given in (6), R is the patient's reward for being cured after the entire episode, t is the patient's *inconvenience cost* associated with each visit, and θ is the imputed cost associated with waiting. Here, we assume that the waiting time will not cause adverse effects (i.e., worsening patient symptoms). This assumption is reasonable for elective surgeries and supported by the empirical observation established by Hurst and Siciliani (2003).

Knowing the revisit rate and waiting time, a patient will seek the service from the public healthcare system only if her expected utility is nonnegative (the outside option for the patient is normalized to zero), i.e., $U(\lambda, \mu) \geq 0$. Here, an implicit assumption is that the patients know the revisit rate and waiting time. Actually, in many countries, such as Australia, both the revisit rate and waiting time are common knowledge (e.g., Australian Hospital Statistics Report 2015). Furthermore, note that when the imputed disutility associated with

each visit or the waiting cost is large enough such that $R-\theta\cdot T(0,\mu)-t\cdot n(\mu)<0$ for all $\mu>0$, the patients' utility is always negative and therefore no patient will seek the treatment from the HCP. To avoid this trivial case, hereafter we assume that $\max_{\mu>0}\{R-\theta\cdot T(0,\mu)-t\cdot n(\mu)\}>0$ so that some patients will seek treatments from the HCP in equilibrium.

By using the fact that the utility $U(\lambda,\,\mu)$ given in (7) is strictly decreasing in λ, we can determine the initial visit rate $\tilde{\lambda}(\mu)$ when visits are endogenously decided by the patients. First, consider the case when the potential initial visit rate Λ (i.e., the potential arrival rate of newly admitted patients) is sufficiently large so that $U(\Lambda,\mu)<0$ (because $U(\lambda,\,\mu)$ is strictly decreasing in λ). In this case, the initial visit rate $\tilde{\lambda}$ in equilibrium satisfies $U(\tilde{\lambda},\mu)=0$, where $\tilde{\lambda}<\Lambda$, and the balking rate equals $(\Lambda-\tilde{\lambda})$ (Hassin and Haviv 2003). We shall refer to this case as the *partial coverage* scenario. Next, consider the case when the potential initial visit rate Λ is sufficiently small so that $U(\Lambda,\mu)\geq 0$. In this case, all potential patients will seek the HCP service so that $\tilde{\lambda}=\Lambda$. We shall refer to this case as the *full coverage* scenario. To avoid repetition and for ease of exposition, we shall present our analysis for the cases where potential patients are either partially covered or fully covered under both schemes in the main text. The analysis for the case where patients are fully covered under one scheme but partially covered under the other scheme is quite similar, which is omitted here. We refer the interested reader to Guo et al. (2017) for details.

So far, we have established the relationships among the initial visit rate λ, the revisit rate $\delta(\mu)$, the patient utility $U(.,.)$, the patient waiting time per visit $W(.,.)$, and the total waiting time $T(.,.)$. Next, we are going to use these relationships to analyze the three-stage Stackelberg game that involves the funder, the HCP and the patients. In this game, the funder (such as the government or a private insurer) first selects the payment scheme (FFS or BP) and the reimbursement rate. Given the reimbursement rate and anticipating the patients' reaction, the HCP determines its service rate μ. Finally, given the service rate μ, the patients decide whether to seek elective treatments from the HCP or not (i.e., patients may balk).

4 Reimbursement Schemes under Partial Coverage: FFS and BP

In this section, we consider the case in which potential patients are *partially covered*. This case is commonly observed in many overcrowded public healthcare systems with long waiting times. We shall use backward induction to analyze the three-stage Stackelberg game for each payment scheme. First, each patient will decide whether or not to seek a visit based on her expected utility. Anticipating the patients' initial visit rate and effective visit rate in

equilibrium, $\tilde{\lambda}(\mu)$ and $\tilde{\lambda}_e(\mu)$, we shall derive the HCP's service rate decision and the funder's reimbursement decisions under both the FFS and BP schemes, respectively. Specifically, for each scheme s, $s = f, b$ (where f and b represent the FFS scheme and the BP scheme, respectively), we first determine the HCP's optimal service rate $\tilde{\mu}_s$. Then, by anticipating the HCP's service rate $\tilde{\mu}_s$ and the corresponding visit rates $\tilde{\lambda}(\tilde{\mu}_s)$ and $\tilde{\lambda}_e(\tilde{\mu}_s)$, we determine the funder's optimal reimbursement rate $\tilde{r}_s$. Table 1 describes the decision sequences under the FFS and BP schemes. It is worth noting that under the FFS scheme, the HCP receives the payment for different types of treatments it provides during each visit. However, as our model focuses on a single type of treatment, we can treat the FFS payment as the payment per visit.

Table 1: Reimbursement Schemes and Sequence of Decisions

Fee-for-Service (FFS)	Bundled Payment (BP)
1. The funder determines the optimal reimbursement rate r_f for each visit.	1. The funder determines the optimal reimbursement rate r_b for each episode.
2. The HCP determines the optimal service rate μ_f.	2. The HCP determines the optimal service rate μ_b.
3. The patients decide to join or to balk, and the initial visit rate in equilibrium is equal to $\tilde{\lambda}(\mu_f)$.	3. The patients decide to join or to balk, and the initial visit rate in equilibrium is equal to $\tilde{\lambda}(\mu_b)$.

4.1 Patients' Joining Decision under Partial Coverage

Under the partial coverage scenario, it is well known that the patient utility in equilibrium equals zero (Hassin and Haviv 2003). By considering the utility function given in (7) and solving $U(\tilde{\lambda}, \mu) = 0$, we can obtain the initial visit rate $\tilde{\lambda}$ in equilibrium under the partial coverage case as

$$\tilde{\lambda}(\mu) = o(\mu) - \frac{\theta(1 - \delta(\mu))}{R(1 - \delta(\mu)) - t}. \tag{8}$$

Note that $\tilde{\lambda}(\mu)$ is decreasing in t, the disutility associated with each visit, which is intuitive.

By accounting for the number of visits over the entire episode $n(\mu)$ given in (4), we can use (3) and (8) to obtain the effective visit rate $\tilde{\lambda}_e$ in equilibrium as follows:

$$\tilde{\lambda}_e(\mu) = n(\mu) \cdot \tilde{\lambda}(\mu) = \mu - \frac{\theta}{R(1 - \delta(\mu)) - t}. \tag{9}$$

It is worth noting from (9) that to ensure the stability of the system (i.e., $\tilde{\lambda}_e(\mu) \leq \mu$), the optimal service rate selected by the HCP should satisfy $R(1 - \delta(\mu)) > t$.

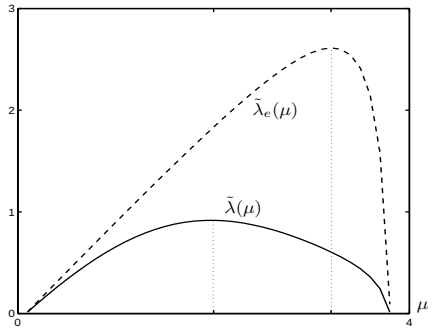

Figure 2: Initial and effective visit rates $\left(\delta(\mu) = \frac{1}{1+e^{-\mu+2}},\ \theta = 0.5,\ t = 1,\ R = 8\right)$

In the next section, we shall utilize the initial visit rate $\tilde{\lambda}(\mu)$ and the effective visit rate $\tilde{\lambda}_e(\mu)$ in equilibrium to derive the HCP's optimal service rate under different payment schemes. In preparation, let us differentiate $\tilde{\lambda}(\mu)$ and $\tilde{\lambda}_e(\mu)$ given in (8) and (9) with respect to μ, getting:

Corollary 1. *Under the partial coverage scenario (i.e., $U(\Lambda, \mu) < 0$), both the initial visit rate $\tilde{\lambda}(\mu)$ and the effective visit rate $\tilde{\lambda}_e(\mu)$ in equilibrium are unimodal in μ. Moreover, the mode of $\tilde{\lambda}(\mu)$ is smaller than that of $\tilde{\lambda}_e(\mu)$; that is, $argmax\{\tilde{\lambda}(\mu)\} \leq argmax\{\tilde{\lambda}_e(\mu)\}$.*

Corollary 1 is induced by the two opposite effects caused by the service rate μ. On one hand, a higher service rate enables the HCP to treat more patients per unit time, which may reduce the waiting time and encourage more patients to seek visits (Rabin 2014). On the other hand, a higher service rate will cause a higher revisit rate, which discourages patients from seeking visits (Varkevisser et al. 2012). Because the cure rate $(1 - \delta(\mu))$ is log-concave, it is less sensitive to the change in μ when μ is small than when it is large. Therefore, when μ is small (large, respectively), the first (second, respectively) effect dominates such that $\tilde{\lambda}(\mu)$ and $\tilde{\lambda}_e(\mu)$ are increasing (decreasing, respectively) in μ. Therefore, $\tilde{\lambda}(\mu)$ and $\tilde{\lambda}_e(\mu)$ are unimodal in μ, as depicted in Figure 2.

Corollary 1 (along with Figure 2) has two implications. First, when μ is large, having the physicians work faster can discourage patients from seeking visits (i.e., both $\tilde{\lambda}(\mu)$ and $\tilde{\lambda}_e(\mu)$ will decrease) because the revisit rate is too high. Second, when μ is moderate, as the mode of $\tilde{\lambda}(\mu)$ is smaller than that of $\tilde{\lambda}_e(\mu)$, having the physicians work faster can reduce the initial visit rate $\tilde{\lambda}(\mu)$ but will increase the effective visit rate $\tilde{\lambda}_e(\mu)$ (due to the significant increase in the revisit rate $\delta(\mu)$). Therefore, when choosing the service rate μ under different payment schemes, our results imply that the HCP may wish to consider the dynamics between the initial visits and followup visits. We shall consider this issue in the next section.

4.2 The HCP's Service Rate Decision under Partial Coverage

Given any reimbursement rate r_s, $s \in \{f, b\}$ and anticipating the equilibrium initial visit rate $\tilde{\lambda}(\mu)$ and the equilibrium effective visit rate $\tilde{\lambda}_e(\mu)$ as given in (8) and (9), the HCP needs to determine its service rate μ_s to maximize its expected profit that is composed of two components: (a) the total amount of reimbursement received from the funder; and (b) the variable cost associated with each patient. First, recall that the HCP is paid r_f for each visit under the FFS scheme and r_b for each episode under the BP scheme. Thus, the HCP receives $r_f\tilde{\lambda}_e(\mu)$ under the FFS scheme and $r_b\tilde{\lambda}(\mu)$ under the BP scheme from the funder. Second, recall that the capacity (i.e., the number of doctors) under our setting is fixed. Hence, the variable cost (i.e., personnel, nurses, consumable items, etc.) is mainly attributed to the length of patients' outpatient visits so that the variable cost per patient visit is $c \cdot (1/\mu)$, where $1/\mu$ is the mean length of outpatient visit and c is the corresponding unit time variable cost incurred by the HCP for treating the patient. Under the BP scheme, the HCP's variable cost associated with each episode of care is $c \cdot (1/\mu) \cdot n(\mu) = c/o(\mu)$ (from (1) and (4)). Then, according to Lemma 1, the HCP's variable cost per episode is first decreasing and then increasing in μ, implying that the HCP faces a tradeoff between a longer service time and the potential savings from the expected readmission reduction (Carey 2015). Combining these observations, we can formulate the HCP's problem under the two schemes as:

$$(FFS) \quad \max_{\mu} \Pi_f(\mu) = r_f \cdot \tilde{\lambda}_e(\mu) - c \cdot \frac{1}{\mu} \cdot \tilde{\lambda}_e(\mu) = \left(r_f - \frac{c}{\mu}\right) \cdot \tilde{\lambda}_e(\mu); \tag{10}$$

$$(BP) \quad \max_{\mu} \Pi_b(\mu) = r_b \cdot \tilde{\lambda}(\mu) - c \cdot \frac{1}{\mu} \cdot \tilde{\lambda}_e(\mu) = \left(r_b - \frac{c}{o(\mu)}\right) \cdot \tilde{\lambda}(\mu). \tag{11}$$

Note that the impact of waiting time and the revisit rate on the HCP's profits $\Pi_f(\mu)$ under the FFS scheme and $\Pi_b(\mu)$ under the BP scheme is indirectly captured by the equilibrium arrival rates $\tilde{\lambda}_e(\mu)$ and $\tilde{\lambda}(\mu)$, respectively. By considering the first-order conditions, we get the following results.

Proposition 1. *For any given reimbursement rate r_s, $s \in \{f, b\}$, the HCP's expected profit $\Pi_s(\mu)$ is unimodal in the service rate μ. In equilibrium,*

1. *the optimal service rate $\tilde{\mu}_f(r_f)$ under the FFS scheme is the unique solution that solves*

$$\frac{d \log \tilde{\lambda}_e(\mu)}{d\mu} = \frac{c}{c\mu - r_f\mu^2}, \tag{12}$$

and $\tilde{\mu}_f(r_f)$ must be larger than the mode of $\tilde{\lambda}_e(\mu)$.

2. *the optimal service rate* $\tilde{\mu}_b(\tilde{r}_b)$ *under the BP scheme is the unique solution that solves*

$$\frac{d\log\tilde{\lambda}(\mu)}{d\mu} = \frac{c\cdot o^{'}(\mu)}{o(\mu)\cdot(c-r_b\cdot o(\mu))}, \tag{13}$$

and $\tilde{\mu}_b(r_b)$ *must be larger than the mode of* $\tilde{\lambda}(\mu)$ *and smaller than* μ^o*, the service rate that maximizes* $o(\mu)$*, that is,* $\tilde{\mu}_b(r_b) < \mu^o$*. Furthermore,* $\tilde{\lambda}(r_b) = \tilde{\lambda}(\tilde{\mu}_b(r_b)) > \tilde{\lambda}(\mu^o)$*.*

First, observe that the unimodality of $\Pi_s(\mu)$ for $s = f, b$ follows immediately from the unimodality of $\tilde{\lambda}_e(\mu)$ and $\tilde{\lambda}(\mu)$, as stated in Corollary 1. Second, by replacing μ with $\tilde{\mu}_s(r_s)$, we can express the aforementioned performance measurements such as revisit rate, visit rate and waiting time as functions of the reimbursement rate r_s. We shall use these measurements to determine the funder's reimbursement decision r_s, $s = f, b$, under the two schemes. In preparation, we first examine the properties of these measurements with respect to the reimbursement rate r_s.

Corollary 2. *The reimbursement rate* r_s*,* $s \in \{f, b\}$*, has the following impact on the following quantities:*

1. *The HCP's optimal service rate* $\tilde{\mu}_s(r_s)$ *and the corresponding revisit rate* $\delta(r_s)$ *are decreasing in* r_s*.*

2. *The initial visit rate* $\tilde{\lambda}(r_s)$*, the waiting time per visit* $W(r_s)$*, the total waiting time* $T(r_s)$*, and the HCP's profit* $\Pi_s(r_s)$ *are increasing in* r_s*.*

The first statement of Corollary 2 reveals that under both the FFS and BP schemes, when the funder offers a higher reimbursement rate r_s, the HCP will set a lower service rate so that physicians spend more time treating each patient. Consequently, fewer patients will revisit the system. To explain this result intuitively, with a higher reimbursement rate r_s, the HCP has a stronger desire to attract more patients to seek visits (i.e., to increase the effective visit rate $\tilde{\lambda}_e(\mu)$ under the FFS scheme and to increase the initial visit rate $\tilde{\lambda}(\mu)$ under the BP scheme). Note that in equilibrium, the HCP selects the service rate that assures a positive marginal profit. Therefore, the right-hand sides of (12) and (13) are negative. Thus, in order to increase $\tilde{\lambda}_e(\mu)$/ $\tilde{\lambda}(\mu)$, the HCP has to lower its service rate μ.

The second statement informs us that a higher reimbursement rate will encourage more patients to seek visits (i.e., the initial visit rate $\tilde{\lambda}(r_s)$ is increasing in r_s), thereby increasing the congestion level of the healthcare system (i.e., both the waiting time per visit $W(r_s)$ and the total waiting time $T(r_s)$ are increasing in r_s). This result can be explained as follows. Recall from above that with a higher reimbursement rate r_s, the HCP will reduce its service rate so as to increase the effective visit rate $\tilde{\lambda}_e(\mu)$ under FFS and increase the initial visit

rate $\tilde{\lambda}(\mu)$ under BP. In addition, with a lower service rate and a higher effective visit rate $\tilde{\lambda}_e(\mu)$ under FFS, it is easy to check from (3) that the initial visit rate $\tilde{\lambda}(\mu)$ under FFS is also higher. Therefore, when the funder offers a higher reimbursement rate r_s, the HCP will lower its service rate so that the initial visit rate and the effective visit rate become higher under both schemes. Consequently, the waiting time per visit and the total waiting time will increase. However, the HCP will earn a higher profit due to the higher visit rate as well as the higher reimbursement rate under both schemes.

In summary, Corollary 2 highlights the trade-off between the revisit rate and the waiting time that the funder needs to strike a balance when considering the reimbursement rate, which we analyze next.

4.3 The Funder's Reimbursement Decision under Partial Coverage

Anticipating the HCP's service rate $\tilde{\mu}_s(r_s)$, $s \in \{f, b\}$ given in (12) and (13) respectively, we now turn our attention to the funder's decision regarding the reimbursement rate r_s. Essentially, the funder selects r_s to maximize the social welfare $S(r_s)$ that takes both the patient utility and the service accessibility into account. Akin to Andritsos and Tang (2014), we model the social welfare by considering the following two factors: (a) the utility $U(.)$ associated with the patient who joins the HCP and receives the treatment (i.e., with public access) and (b) an additional cost β incurred by the public system for each patient who balks due to long waiting times (i.e., without public access). Recall that $\tilde{\lambda}(r_s)$ is the initial visit rate and $(\Lambda - \tilde{\lambda}(r_s))$ is the balking rate. By noting that each admitted patient obtains a utility $U(\tilde{\lambda}(\mu), \mu)$ and each balking patient incurs an additional cost β, the social welfare can be formulated as

$$S(r_s) = \tilde{\lambda}(r_s) \cdot U(\tilde{\lambda}(r_s), \tilde{\mu}_s(r_s)) - \beta(\Lambda - \tilde{\lambda}(r_s)). \tag{14}$$

Note that different from the customer welfare definition in the classic queueing literature (e.g., Hassin and Haviv 2003) which often considers the customer welfare as the sum of customer utilities, the funder's objective function here includes an additional cost associated with the balking patients. The additional cost term $\beta(\Lambda - \tilde{\lambda}(r_s))$ can be justified and interpreted as follows. First, the health policies that fail to meet the needs of the vulnerable people are regarded as a violation of human rights (Fleck 2002). Therefore, the funder incurs a *political cost* for failing to delivery the healthcare service to those balking patients (i.e., $\Lambda - \tilde{\lambda}(r_s)$). Second, the balking patients have to seek the healthcare service elsewhere in the outside systems, which brings burden to those systems and imposes negative externalities

on the patients there. Such negative impact is captured by the negative term in the welfare function (14).

Furthermore, due to limited financial resources, the parliament/council in many countries will determine the earmarked budget for public healthcare in advance. For example, in 2016 the National Health Service in England earmarked an extra £2.4 billion a year for general practice services by 2020/21 (General Practice Forward View 2016). Therefore, we assume that the funder is subject to a limited budget B. By taking the funder's budget B and the HCP's participation constraint into consideration, we can formulate the funder's problem under partial coverage as follows:

$$\max_{r_s} S(r_s) = U(\tilde{\lambda}(r_s), \tilde{\mu}_s(r_s)) \cdot \tilde{\lambda}(r_s) - \beta \cdot (\Lambda - \tilde{\lambda}(r_s)) = -\beta(\Lambda - \tilde{\lambda}(r_s)), \tag{15}$$

$$\begin{aligned} s.t. \qquad & Budget\ constraint : \begin{cases} (FFS)\ r_f \cdot \tilde{\lambda}_e(r_f) \leq B, \\ (BP)\ r_b \cdot \tilde{\lambda}(r_b) \leq B, \end{cases} \\ & Participation\ constraint : \Pi_s(r_s) \geq 0. \end{aligned} \tag{16}$$

Recall from §4.1 that $U(\tilde{\lambda}(r_s), \tilde{\mu}_s(r_s)) = 0$ under the partial coverage case. Thus, it is easy to check from (15) that the funder's objective is equivalent to maximizing the accessibility (i.e., the initial visit rate $\tilde{\lambda}(r_s)$). This is consistent with the practice that, in many public healthcare systems that serve a large patient population, the government's overarching objective is to improve accessibility by maximizing $\tilde{\lambda}(r_s)$ (Aboolian et al. 2016). Combining this observation and the fact that $\tilde{\lambda}(r_s)$ is increasing in r_s (Corollary 2), we obtain the following result.

Proposition 2. *The social welfare $S(r_s)$ given in (15) is increasing in r_s. The funder's budget constraint (16) is binding so that the funder's optimal reimbursement rate $\tilde{r}_s$ satisfies $\tilde{r}_f \cdot \tilde{\lambda}_e(\tilde{r}_f) = B$ under the FFS scheme and $\tilde{r}_b \cdot \tilde{\lambda}(\tilde{r}_b) = B$ under the BP scheme.*

Proposition 2 implies that when serving a large patient population (e.g., the UK), the public healthcare system can only provide partial coverage, and the funder always exhausts its budget under both the FFS and BP schemes (Donnelly 2013). We can further establish the following comparative statistics with respect to the budget B.

Corollary 3. *Under both the FFS and BP schemes, the HCP's optimal service rate $\tilde{\mu}_s(\tilde{r}_s)$ and the corresponding revisit rate $\delta(\tilde{r}_s)$ are decreasing in B. However, the optimal reimbursement rate $\tilde{r}_s$, the initial visit rate $\tilde{\lambda}(\tilde{r}_s)$, the waiting time per visit $W(\tilde{r}_s)$, and the total waiting time $T(\tilde{r}_s)$ are increasing in B.*

Corollary 3 reveals that under the partial coverage case, the funder can afford to offer a higher reimbursement rate with a higher budget. By considering this key result, we can

apply Corollary 2 to interpret all other results stated in Corollary 3. To avoid repetition, we omit the details. Hence, under both the FFS and BP schemes, increasing the funder's budget B can improve patient access and reduce the revisit rate, but it can increase the waiting time. This implication will play a role when we compare the performance between the FFS scheme and the BP scheme.

4.4 Performance Comparison under Partial Coverage: FFS vs. BP

So far, we have derived the equilibrium outcomes of different performance metrics (the social welfare, initial visit rate, service rate, revisit rate and waiting times) associated with both payment schemes under the partial coverage scenario, as presented in Propositions 1 and 2. We now evaluate the performance of the healthcare system under the FFS scheme and the BP scheme by comparing these performance outcomes.

4.4.1 The Benchmark: A Centralized Healthcare System

Before we compare the performance of the healthcare system under the FFS scheme and the BP scheme, let us first establish a "benchmark" that is based on a centralized healthcare system in which the HCP is fully owned by the funder such that the service rate is determined by the funder. At the same time, there is no explicit payment for the HCP so that the payment scheme is irrelevant in the centralized system. Patients, however, still make their decentralized (independent) joining/balking decisions. This centralized healthcare system fits the setting where the funder and the healthcare provider belong to the same organization.

In the centralized healthcare system, the aforementioned variable cost $c \cdot \tilde{\lambda}_e(\mu)/\mu$ is now borne by the funder; hence, the funder's problem can be formulated as:

$$\max_{\mu} S(\mu) = U(\tilde{\lambda}(\mu)) \cdot \tilde{\lambda}(\mu) - \beta \left(\Lambda - \tilde{\lambda}(\mu)\right) = -\beta \left(\Lambda - \tilde{\lambda}(\mu)\right), \tag{17}$$

$$s.t. \qquad c \cdot \tilde{\lambda}_e(\mu) \cdot \frac{1}{\mu} \leq B. \tag{18}$$

By investigating the first-order condition of $S(\mu)$, we obtain the following results that involve the mode of $\tilde{\lambda}(\mu)$, as examined in Corollary 1. For ease of exposition, let us define this mode as μ_c, where $\mu_c = argmax\{\tilde{\lambda}(\mu)\}$.

Proposition 3. *In a centralized healthcare system, when potential patients are partially covered, the social welfare $S(\mu)$ given in* (17) *is unimodal in μ, and the optimal service rate μ^* can be described as follows:*

1. *If the funder's budget is relatively large so that $B \geq c \cdot \tilde{\lambda}_e(\mu_c)/\mu_c$, then $\mu^* = \mu_c$ and $\mu^* < \mu^o$.*

2. *If the funder's budget is relatively small so that* $B < c \cdot \tilde{\lambda}_e(\mu_c)/\mu_c$*, the budget constraint* (18) *is binding and* μ^* *is the larger service rate that satisfies* $c \cdot \tilde{\lambda}_e(\mu^*)/\mu^* = B$*.*

Proposition 3 can be explained as follows. Recall that the funder's objective under the partial coverage scenario is equivalent to maximizing the accessibility, that is, the initial visit rate $\tilde{\lambda}(\mu)$. Therefore, in the centralized system, it is optimal for the funder to set μ^* equal to μ_c (i.e., the mode of $\tilde{\lambda}(\mu)$) if μ_c is a feasible solution so that $B \geq c \cdot \tilde{\lambda}_e(\mu_c)/\mu_c$. In the event that μ_c is not feasible, then the budget constraint (18) is binding. In this case, it is optimal for the funder to set μ^* that keeps the budget constraint binding. Based on the fact from Corollary 1 that $\tilde{\lambda}_e(\mu)$ is unimodal in μ, it is easy to check that the variable cost $c \cdot \tilde{\lambda}_e(\mu)/\mu$ is also unimodal in μ. This observation implies that there are two service rates that keep the budget constraint binding. Combining this observation with the fact that the funder aims to maximize accessibility, the funder will choose the larger root that solves $c \cdot \tilde{\lambda}_e(\mu^*)/\mu^* = B$.

4.4.2 Performance Comparison

Using the benchmark established in Proposition 3, we now evaluate the performance of the FFS scheme and the BP scheme. Specifically, our direct comparison yields the following results.

Proposition 4. *Consider the case in which potential patients are partially covered under both the FFS and BP schemes. Then, in comparison, we establish the following results:*

1. *Both the social welfare and the initial visit rate under the BP scheme are higher than those under the FFS scheme but are lower than those under the centralized healthcare system, that is,* $S(\mu^*) > S(\tilde{r}_b) > S(\tilde{r}_f)$ *and* $\tilde{\lambda}(\mu^*) > \tilde{\lambda}(\tilde{r}_b) > \tilde{\lambda}(\tilde{r}_f)$*.*

2. *Both the service rate and the revisit rate under the BP scheme are lower than those under the FFS scheme but are larger than those under the centralized healthcare system, that is,* $\mu^* < \tilde{\mu}_b(\tilde{r}_b) < \tilde{\mu}_f(\tilde{r}_f)$ *and* $\delta(\mu^*) < \delta(\tilde{r}_b) < \delta(\tilde{r}_f)$*.*

3. *Both the waiting time per visit and the total waiting time under the BP scheme are higher than those under the FFS scheme but are lower than those under the centralized healthcare system, that is,* $W(\mu^*) > W(\tilde{r}_b) > W(\tilde{r}_f)$ *and* $T(\mu^*) > T(\tilde{r}_b) > T(\tilde{r}_f)$*.*

Proposition 4 implies that when potential patients are partially covered, the BP scheme dominates the FFS scheme in terms of the social welfare (in terms of higher $S(.)$) and service quality (in terms of both a lower service rate and a lower revisit rate). However, the FFS scheme outperforms the BP scheme in terms of shorter waiting times $W(.)$ and $T(.)$. These results can be explained as follows. Recall that the BP scheme pays the HCP a fixed amount

for each admitted patient no matter how many times the patient revisits the system. Hence, the HCP under the BP scheme has incentives to reduce the service rate so as to reduce the revisit rate. However, reducing the revisit rate results in a higher initial visit rate under the BP scheme. Furthermore, from (7), if the potential patients are partially covered, then the initial visit rate solves $U(\tilde{\lambda}, \mu) = 0$. Due to a lower revisit rate, admitted patients under the BP scheme can tolerate a longer waiting time in equilibrium such that $W(\tilde{r}_b) > W(\tilde{r}_f)$ and $T(\tilde{r}_b) > T(\tilde{r}_f)$. As maximizing the social welfare is equivalent to maximizing the initial visit rate under the partial coverage scenario, the social welfare under the BP scheme is also larger.

Furthermore, Proposition 4 also implies that the performance metrics of the BP scheme are "closer" to the benchmark case (associated with the centralized system) than those of the FFS scheme. In particular, both the total waiting time and the waiting time per visit under the centralized system are the longest in relation to those under the FFS and BP schemes. This is because in the centralized system the funder has direct control over the service rate. Therefore, as the funder's objective is to maximize accessibility (i.e., social welfare $S(.)$), it is more willing to sacrifice the waiting times.

Finally, the results stated in Proposition 4 are consistent with the findings of the previous studies that show the FFS scheme is effective in reducing the waiting time but not in improving the service quality. For example, Blomqvist and Busby (2013) show that the FFS scheme is effective in reducing the waiting time in Canada. Mot (2002) finds that in the Netherlands, the abolition of the FFS scheme has caused the waiting time to increase for elective surgery. These two results corroborate the third statement of Proposition 4.

In summary, in considering the partial coverage case that occurs when the patient population is large, we find that there is no dominant scheme. The BP scheme is more effective in improving the social welfare and reducing the revisit rate; however, the FFS scheme is more effective in reducing the waiting time. Next, we examine the full coverage case that occurs when the patient population is small. As we shall see, the results as stated in Propositions 2 and 4 no longer hold.

5 Reimbursement Schemes under Full Coverage: FFS and BP

We now consider the case when the potential patients are fully covered under both the FFS and BP schemes. This case is suitable for public HCPs in rural, which normally have low patient volumes (DiChiara 2015). Recall that under the full coverage scenario, $U(\Lambda, \mu) \geq 0$ (see §3.3). This condition can be further simplified as $\tilde{\lambda}(\mu) \geq \Lambda$ after some algebra. With a

little abuse of notation, we use $\tilde{\lambda}(\mu)$ here to represent the expression of the right-hand side of (8).

Under the full coverage scenario, the initial visit rate is given by Λ and the effective visit rate is $\Lambda/(1-\delta(\mu))$. By noting that the average variable cost associated with each patient's visit is $c \cdot (1/\mu)$, we can formulate the HCP's problem under the full coverage scenario as follows:

$$\begin{aligned} (FFS)\ \max_{\mu}\ \Pi_f(\mu) &= \left(r_f - \frac{c}{\mu}\right)\frac{\Lambda}{1-\delta(\mu)}, && (19)\\ s.t. \quad & \tilde{\lambda}(\mu) \geq \Lambda; \\ (BP)\ \max_{\mu}\ \Pi_b(\mu) &= \left(r_b - \frac{c}{o(\mu)}\right)\Lambda, && (20)\\ s.t. \quad & \tilde{\lambda}(\mu) \geq \Lambda, \end{aligned}$$

where the constraint $\tilde{\lambda}(\mu) \geq \Lambda$ guarantees that the healthcare system offers full coverage to all potential patients.

Proposition 5. *Consider the case in which all potential patients are fully covered under both the FFS and BP schemes. Then, for any given reimbursement rate r_s, $s \in \{f, b\}$,*

1. *the HCP's optimal service rate under the FFS scheme is the largest μ satisfying $\tilde{\lambda}(\mu) \geq \Lambda$ (i.e., $\tilde{\mu}_f = \max\{\mu,$ subject to $\tilde{\lambda}(\mu) \geq \Lambda\}$), where $\tilde{\lambda}(\mu)$ is given as in (8).*

2. *the HCP's optimal service rate under the BP scheme satisfies*

$$\tilde{\mu}_b = \begin{cases} \mu^o, & \text{if } \tilde{\lambda}(\mu^o) \geq \Lambda, \\ \max\{\mu : \tilde{\lambda}(\mu) \geq \Lambda\} & \text{if } \tilde{\lambda}(\mu^o) < \Lambda. \end{cases}$$

Under both schemes, the optimal service rate $\tilde{\mu}_s$, $s = f, b$, is independent of the funder's reimbursement rate r_s.

The intuition behind Proposition 5 is as follows. First, observe from (19) that the HCP's variable cost c/μ decreases in μ while its effective arrival rate $\Lambda/(1-\delta(\mu))$ increases in μ. Thus, the HCP's profit $\Pi_f(\mu)$ under the FFS scheme is always increasing in μ. These observations imply that under the FFS scheme, the HCP will choose the largest service rate in equilibrium that ensures the potential patients are fully covered (i.e., $\max\{\mu : \tilde{\lambda}(\mu) \geq \Lambda\}$). Our results reveal that when the patient population is small, the FFS scheme creates an incentive for the HCP to increase its service rate so as to generate as much revenue as possible.

Next, under the BP scheme, observe from (20) that the HCP's objective is equivalent to minimizing the variable cost $c/o(\mu)$. According to Lemma 1, the cure service rate $o(\mu)$

is unimodal in μ and reaches its maximum at μ^o. Therefore, $\Pi_b(\mu)$ given in (20) is also unimodal in μ and its corresponding mode is also μ^o. When μ^o is feasible under the full coverage scenario (i.e., $\tilde{\lambda}(\mu^o) \geq \Lambda$), it is natural that the HCP will choose μ^o in equilibrium. However, when μ^o is infeasible under the full coverage scenario (i.e., $\tilde{\lambda}(\mu^o) < \Lambda$), the HCP will choose the largest service rate that ensures the potential patients are fully covered (i.e., $\max\{\mu : \tilde{\lambda}(\mu) \geq \Lambda\}$).

Finally, unlike the partial coverage scenario, Proposition 5 implies that under the full coverage scenario, the HCP's optimal service rate is independent of the funder's reimbursement rate. Because the funder cannot regulate the HCP's service rate decision under both schemes, the funder will select the smallest feasible reimbursement rate. Because all the performance metrics that we are going to compare, such as the social welfare, the initial visit rate and the total waiting time, only depend on the service rate and the initial visit rate, for ease of exposition we omit the analysis of the funder's reimbursement rate decisions under the full coverage case. (We refer the readers to the online Appendix B of Guo et al. (2017) for details.)

5.1 Performance Comparison

Based on the equilibrium outcomes under the full coverage scenario as established in Proposition 5, we now compare the performance of the healthcare system under the FFS and BP schemes. To this end, similar to the partial coverage scenario, we first consider a centralized healthcare system in which the HCP is fully owned by the funder as the benchmark.

5.1.1 The Benchmark: A Centralized Healthcare System

Because the HCP is fully owned by the funder in a centralized healthcare system, the service rate is determined by the funder and the aforementioned variable cost $c \cdot \tilde{\lambda}_e(\mu)/\mu$ is also borne by the funder. By noting that under the full coverage scenario, the patient's utility is positive and there is no penalty cost from the balking patients, the funder's problem can be formulated as:

$$\max_{\mu} S(\mu) = \Lambda \cdot U(\Lambda, \mu); \tag{21}$$

$$s.t. \quad c \cdot \Lambda \cdot \frac{1}{o(\mu)} \leq B, \tag{22}$$

$$\tilde{\lambda}(\mu) \geq \Lambda.$$

It is worth noting that when the maximum cure service rate $o(\mu^o)$ is so small such that $o(\mu^o) < c \cdot \Lambda/B$, the funder's budget cannot cover the variable cost under the full coverage

scenario, which further implies that the healthcare system cannot achieve the full coverage scenario in equilibrium. In view of this, we assume $o(\mu^o) \geq c \cdot \Lambda/B$ hereafter.

Proposition 6. *In a centralized healthcare system, when potential patients are fully covered, both the social welfare $S(\mu)$ given in (21) and the patient utility $U(\Lambda, \mu)$ are unimodal in μ. The funder's optimal service rate μ^* can be described as follows:*

1. *If the funder's budget is relatively large so that $B \geq c \cdot \Lambda/o(\mu_u)$, then $\mu^* = \mu_u$, where $\mu_u = argmax\{U(\Lambda, \mu)\}$ and $\mu^* < \mu^o$.*

2. *If the funder's budget is relatively small so that $B < c \cdot \Lambda/o(\mu_u)$, the budget constraint (22) is binding and μ^* is the smaller service rate that satisfies $o(\mu^*) = c \cdot \Lambda/B$.*

Proposition 6 can be explained as follows. A close look of (21) shows that under the full coverage scenario, the funder's objective is equivalent to maximizing the patient utility. Therefore, in the centralized system, it is optimal for the funder to set μ^* equal to μ_u (i.e., the service rate that maximizes the patient utility) if μ_u is a feasible solution that satisfies $B \geq c \cdot \Lambda/o(\mu_u)$. In the event that μ_u is not feasible, then the budget constraint (22) is binding; that is, $o(\mu) = c \cdot \Lambda/B$. Because the cure service rate $o(\mu)$ is unimodal in μ (Lemma 1), there are two roots that solve $o(\mu) = c \cdot \Lambda/B$. As $\delta(\mu)$ is increasing in μ, from (7) we can easily know that the patient utility $U(\Lambda, \mu)$ is larger at the smaller root. Combining this observation with the fact that the funder aims to maximize the patient utility, the funder will choose the smaller root that solves $o(\mu^*) = c \cdot \Lambda/B$.

5.1.2 Performance Comparison

So far, we have derived the equilibrium outcomes associated with the FFS scheme, the BP scheme and the centralized healthcare system when potential patients are fully covered. We now evaluate the performance of the healthcare system under the FFS scheme and the BP scheme. To guarantee that the healthcare system in equilibrium achieves the full coverage under both the FFS and BP schemes, the equilibrium outcome under the partial coverage must be infeasible; that is, $\Lambda \leq \tilde{\lambda}(\tilde{r}_s)$, $s \in \{f, b\}$, where $\tilde{\lambda}(\tilde{r}_s)$ represents the initial visit rate under the partial coverage case. According to Proposition 4, this condition is equivalent to $\Lambda \leq \tilde{\lambda}(\tilde{r}_f)$.

Corollary 4. *Suppose that potential patients are fully covered under both the FFS and BP schemes (i.e., when $\Lambda \leq \tilde{\lambda}(\tilde{r}_f)$). Then,*

1. *If the potential patient population Λ is medium so that $\tilde{\lambda}(\mu^o) \leq \Lambda \leq \tilde{\lambda}(\tilde{r}_f)$, the optimal service rates under the FFS and BP schemes are equal, which is larger than the optimal*

service rate under the centralized healthcare system; that is, $\tilde{\mu}_f = \tilde{\mu}_b = \max\{\mu : \tilde{\lambda}(\mu) \geq \Lambda\} > \mu^*$. *Consequently, the social welfare, the revisit rate, the initial visit rate, the waiting time per visit and the total waiting time are the same under both the FFS and BP schemes.*

2. *If the potential patient population* Λ *is so small such that* $\Lambda < \min\{\tilde{\lambda}(\mu^o), \tilde{\lambda}(\tilde{r}_f)\}$, *then the BP scheme dominates the FFS scheme in terms of the social welfare, service quality and the total waiting time; that is,* $S(\tilde{\mu}_f) < S(\tilde{\mu}_b)$, $\delta(\tilde{\mu}_f) > \delta(\tilde{\mu}_b)$ *and* $T(\tilde{\mu}_f) > T(\tilde{\mu}_b)$.

3. *Relative to the FFS and BP schemes, under the centralized healthcare system the social welfare is larger while both the service rate and the revisit rate are smaller. Furthermore, the total waiting time under the centralized healthcare system is larger than that under the BP scheme.*

The first statement of Corollary 4 reveals that the FFS and BP schemes are equally efficient if the patient population is medium (i.e., $\tilde{\lambda}(\mu^o) \leq \Lambda \leq \tilde{\lambda}(\tilde{r}_f)$). However, when the patient population is very small (i.e., $\Lambda < \min\{\tilde{\lambda}(\mu^o), \tilde{\lambda}(\tilde{r}_f)\}$), the BP scheme dominates the FFS scheme in terms of the social welfare, service quality and congestion level. These results can be explained as follows. When $\tilde{\lambda}(\mu^o) \leq \Lambda \leq \tilde{\lambda}(\tilde{r}_f)$, Proposition 5 reveals that the HCP will select the largest service rate that makes the potential patients fully covered under both the FFS and BP schemes so that $\tilde{\mu}_f = \tilde{\mu}_b$. Therefore, the FFS and BP schemes are equally efficient in terms of all the performance metrics.

Next, when $\Lambda < \min\{\tilde{\lambda}(\mu^o), \tilde{\lambda}(\tilde{r}_f)\}$, μ^o is feasible under the full coverage and therefore, $\tilde{\mu}_b = \mu^o$. Because the optimal service rate under the FFS scheme $\tilde{\mu}_f$ is the largest one that makes the potential patients fully covered, $\tilde{\mu}_f > \tilde{\mu}_b = \mu^o$. As the revisit rate $\delta(\mu)$ is increasing in μ, $\delta(\tilde{\mu}_f) > \delta(\tilde{\mu}_b)$. From (6) we can easily know that the total waiting time T is decreasing in $o(\mu)$. As $o(\mu)$ is maximized at μ^o, $T(\tilde{\mu}_b) < T(\tilde{\mu}_f)$. Consequently, compared with the FFS scheme, the service quality is better and the total waiting cost is smaller under the BP scheme. Therefore, the social welfare under the BP scheme is also larger than that under the FFS scheme.

We can conclude from the first two statements of Corollary 4 that the BP scheme weakly dominates the FFS scheme when potential patients are fully covered. And the last statement of Corollary 4 implies that the BP scheme is "closer" to the benchmark case (associated with the centralized system) than the FFS scheme in terms of social welfare and the revisit rate. Furthermore, the total waiting time under the BP scheme is smaller than that associated with the centralized system; however, the total waiting time under the centralized system is not necessarily smaller than that under the FFS scheme. This is because under the full coverage scenario, the HCP's objective under the BP scheme is actually to maximize the cure

service rate $o(\mu)$, and according to (6), maximizing $o(\mu)$ is equivalent to minimizing the total waiting time T. However, the funder's objective under the centralized system is equivalent to maximizing the patient utility, and therefore, it faces an intertemporal trade-off between the service quality and the total waiting time. If the cost associated with the low service quality is large (via a higher t, the non-pecuniary disutility associated with each visit), it may sacrifice the total waiting time such that the total waiting time under the centralized system may be larger than that under the FFS scheme.

In summary, we can conclude from Corollary 4 that under the full coverage scenario, when the patient population is medium, both schemes yield the same performance. However, when the patient population is very small, the BP scheme dominates the FFS scheme in terms of the social welfare, service quality and total waiting time.

5.2 Numerical Illustration

To facilitate our exposition, we numerically illustrate the comparison results here. Recall that when $\tilde{\lambda}(\tilde{r}_f) < \Lambda \leq \tilde{\lambda}(\tilde{r}_b)$, the potential patients are fully covered under the BP scheme but are partially covered under the FFS scheme. We also study this case in the online Apendix B. The corresponding comparison results are similar to Proposition 4 and Corollary 4. Specifically, when $\tilde{\lambda}(\mu^o) \leq \tilde{\lambda}(\tilde{r}_f)$, the results given in Proposition 4 still hold: the BP scheme dominates the FFS scheme in terms of the social welfare and service quality but the FFS scheme outperforms the BP scheme in terms of the congestion level. However, when $\tilde{\lambda}(\mu^o) > \tilde{\lambda}(\tilde{r}_f)$, there exists a threshold $\bar{\Lambda} \in [\tilde{\lambda}(\tilde{r}_f), \tilde{\lambda}(\mu^o)]$ such that when Λ is relatively large (i.e., $\Lambda > \bar{\Lambda}$), the results given in Proposition 4 hold; and when Λ is relatively small (i.e., $\Lambda \leq \bar{\Lambda}$), the results given in the second statement of Corollary 4 hold (i.e., the BP scheme dominates the FFS scheme in terms of the social welfare, service quality and total waiting time). Specifically, we illustrate the case $\tilde{\lambda}(\mu^o) \leq \tilde{\lambda}(\tilde{r}_f)$ in Figure 3: when potential patients are fully covered under both the FFS and BP schemes (the light blue shaded area), Corollary 4 holds; when potential patients are fully covered under the BP scheme but are partially covered under the FFS scheme (the white area), the results given in Corollary 4 still hold; when potential patients are partially covered under both the FFS and BP schemes (the grey shaded area), Proposition 4 holds.

6 Extension: A Blended Payment Scheme

So far, we have shown that both the FFS and BP schemes cannot achieve the social optimum under both the partial coverage scenario and the full coverage scenario (relative to the centralized system). In practice, there are also advocates for using a combination of different

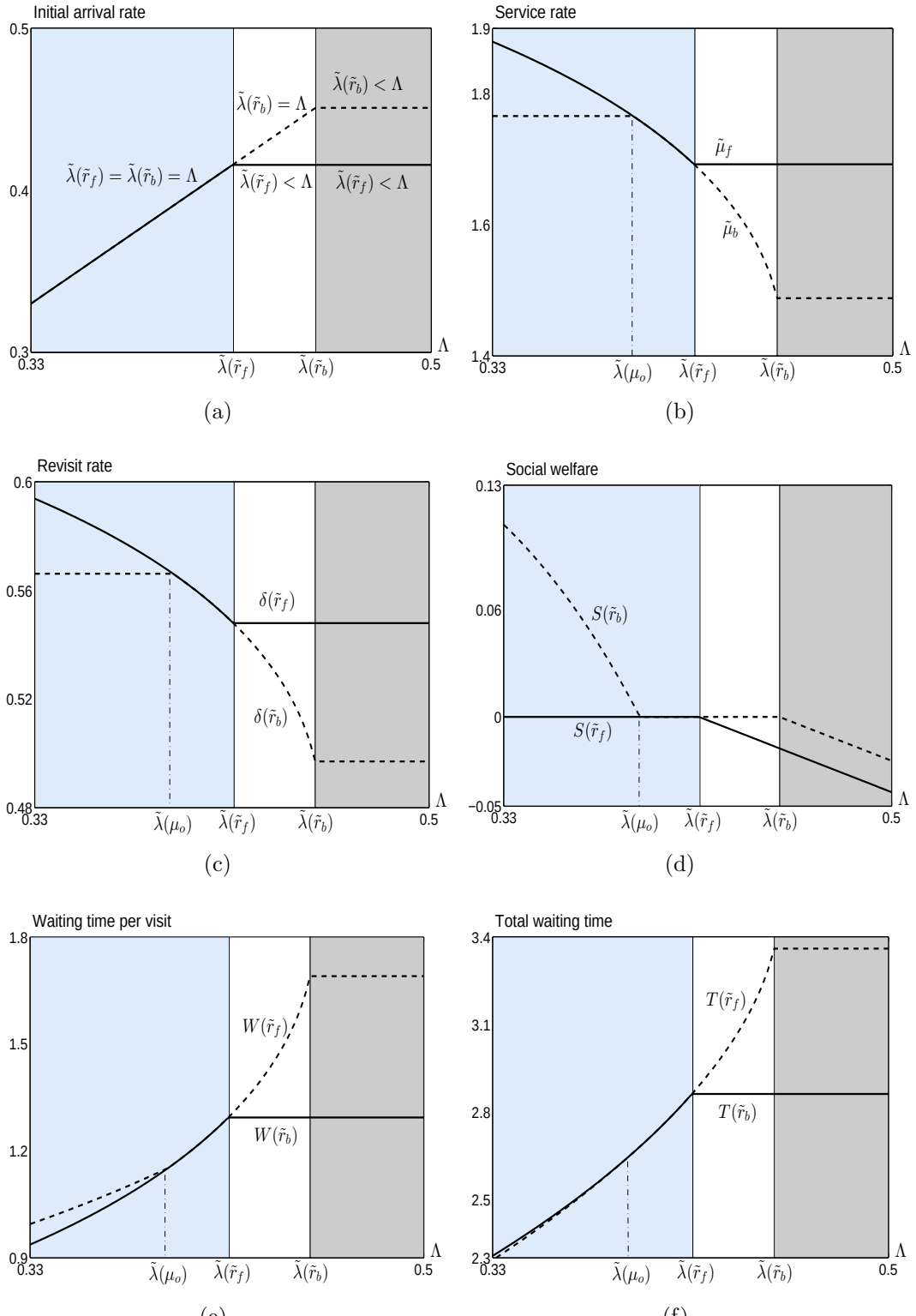

Figure 3: Performance comparison $\big(\delta(\mu) = \frac{1}{1+e^{-\mu+1.5}}$, $\theta = 0.9$, $t = 2$, $R = 7$, $B = 2.5$, $\beta = 0.5$, $c = 1\big)$, where the light blue shaded area represents full coverage under both the FFS and BP schemes; the white area represents full coverage under the BP scheme but partial coverage under the FFS scheme; and the grey shaded area represents partial coverage under both the FFS and BP schemes.

payment schemes. For example, in Australia, to efficiently allocate resources and overcome the shortcomings of the different payment schemes, the primary healthcare advisory group suggests the Australian government design blended payments, that is, a mix of multiple payment schemes including the FFS and BP schemes (Australian Primary Health Care Advisory Group Report 2015). In view of this, we consider a blended payment scheme, which can be characterized by two parameters (r_m, α). Specifically, under the blended payment scheme, the HCP receives a reimbursement rate r_m for the first visit of a patient and obtains a "discounted reimbursement rate" αr_m, $\alpha \in [0, 1]$, for each subsequent visit(s) of the patient. For ease of exposition, we use MP to represent the blended payment scheme. Note that the blended payment scheme is a generalized payment scheme because it reduces to the FFS scheme when $\alpha = 1$ and to the BP scheme when $\alpha = 0$. The main results are summarized in the following proposition. For more details, see Guo et al. (2017).

Proposition 7. *The performance of the MP scheme falls into the middle of those of the FFS and BP schemes in terms of social welfare, the revisit rate and the total waiting time. Moreover, the social welfare is the highest under the BP scheme while the total waiting time is the shortest under the FFS scheme.*

Proposition 7 implies that both the FFS and BP schemes have their own pros and cons. Blending them together can somehow mitigate their respective defects but also diminishes the benefits of each scheme a little. Which scheme should be adopted? The Hong Kong government may find the FFS scheme more attractive, as the waiting time for elective care is extremely long (e.g., the average waiting time for cataract surgery can be two years; see Official Website of Hong Kong Hospital Authority 2016). The government of the United States may find the BP scheme more attractive, as waiting time is not a major concern in its mainly private healthcare system. The Australian government perhaps faces a situation in between that of Hong Kong and the United States and can certainly consider the blended system.

Combining the results stated in Propositions 4 and 7, we can conclude that the MP scheme cannot achieve the social optimum either (relative to the centralized system). It is worth noting that Adida et al. (2017) also consider a hybrid payment system, which is a combination of BP and FFS schemes. However, they obtain the opposite result: when the HCP is risk neutral, the hybrid system can achieve the system optimum. Such difference is caused by the different assumptions in Adida et al. (2017) and our paper. In Adida et al. (2017), waiting time is ignored and the HCP can reject a patient, two assumptions that fit the private care setting well. However, we consider a public care, where long waiting time is common and the HCP cannot cherry-pick patients. Therefore, the patient arrival process

in these two systems are managed in different ways, leading to the different preferences over the BP, MP and FFS payment schemes for the public and private care.

7 Concluding Remarks

In this paper, we present a queueing model with an endogenous arrival rate selected by the patient and an endogenous revisit rate controlled by the HCP (via the selected service rate). By analyzing a three-stage Stackelberg game with an embedded queueing model, we compare the performance associated with the FFS and BP reimbursement schemes. By considering the an intertemporal trade-off between service rate and service quality (in terms of the revisit rate), we have obtained the following managerial insights. First, when potential patients are partially covered, we find that a higher service rate may reduce both the initial visit rate and the effective visit rate under both schemes. Second, we show that under the partial coverage, a higher reimbursement rate can improve the service quality in terms of the revisit rate but can increase the waiting time under both the FFS and BP reimbursement schemes.

More importantly, by investigating the funder's reimbursement decisions and comparing the equilibrium outcomes associated with the two schemes, we find that the BP reimbursement scheme may not always dominate the FFS reimbursement scheme. Specifically, when the potential patients are partially covered, the BP scheme dominates the FFS scheme in terms of the social welfare and service quality (i.e., the revisit rate); however, the FFS scheme outperforms the BP scheme in terms of the waiting time per visit and the total waiting time.

When the potential patients are fully covered, the BP scheme weakly dominates the FFS scheme in terms of social welfare, service quality and the total waiting time. In particular, when the patient population is medium, the FFS and BP schemes are equally efficient in terms of all performance metrics, including the revisit rate, the waiting time per visit, the total waiting time and the social welfare. Overall, the implications of our findings are as follows. First, when the size of the patient population is large, shifting from the FFS scheme to the BP scheme can improve the social welfare and reduce revisits, but it can increase the waiting time. Second, when potential patients are fully covered and the size of the patient population is moderate, the two schemes yield the same outcomes. In such case, it seems unnecessary to move from the FFS scheme to the BP scheme. However, when the size of the patient population is very small, the BP scheme dominates the FFS scheme.

A major finding of this paper is that the BP scheme has a larger benefit in a less-congested system compared to the FFS scheme. This finding can perhaps shed light on the interesting capacity trend observed in the healthcare system of the United States after implementing the BP scheme. It is observed that after the BP scheme is implemented, the HCPs tend

to merge into larger units with more resources and larger capacities (Augenstein and Livio 2012). Coupled with the queueing knowledge that the pooling of multiple service facilities can generally reduce the waiting time for customers, our results might explain the benefit of doing this: the capacity-merging trend can make the HCP benefit more from the BP scheme, as it can somehow avoid the defect of increased patient waiting time associated with the BP scheme.

Our analysis represents an initial attempt to examine the performance of the healthcare reimbursement scheme by capturing the strategic interactions among the patients, the HCP, and the funder and by taking into account the relationship between service quality (in terms of the revisit rate) and service speed. Admittedly, our model has some limitations and the relaxation of certain assumptions is worth further investigation. First, we have assumed that patients have perfect information about the HCP's service quality, and it is of interest to examine a situation when there is information asymmetry between patients and HCPs. Second, we consider elective non-urgent outpatient care in the paper. It would be interesting to consider payment schemes for other types of care, such as urgent care and the inpatient system. Third, patients are assumed to be homogeneous in our work. In the future, one might consider a more differentiated market with different types of patients.

Another important aspect to consider is the competition among different HCPs. In the presence of market competition, the non-cured patients may seek visits from other HCPs. (For example, Andritsos and Tang (2014) examine the impact of competition between the private and public HCPs on the social welfare in Europe.) Therefore, the HCP may not generate more demand by reducing the revisit rate, which can dramatically change the HCPs' choices of service rate, as demonstrated in our paper. Yet other issue is that our model focuses on the public care services that are basically free for the patients. For other healthcare systems involved with insurance companies and patients' co-payment, the game is even more complex as more parties are involved in the game. We leave the analysis of such systems to future research.

Acknowledgments

We are grateful to the department editor (Professor Morris Cohen), an anonymous associate editor, and two anonymous referees for very helpful comments and suggestions. The first author acknowledges the financial support by the Hong Kong RGC GRF (Grant No. PolyU 15504515). The third author acknowledges the financial support by the Research Grants Council of Hong Kong (RGC Reference Number: PolyU 15504615). The fourth author is the corresponding author and was supported by the National Natural Science Foundation of

China (Grant No. 71501160).

References

Aboolian, R., O. Berman, V. Verter. 2016. Maximal accessibility network design in the public sector. *Transportation Science* **50**(1) 336-347.

Adida, E., H. Mamani, S. Nassiri. 2017. Bundled Payment vs. Fee-for-Service: impact of payment scheme on performance. *Management Science* **63**(5) 1606-1624.

Alizamir, S., F. de Véricourt, P. Sun. 2013. Diagnostic accuracy under congestion. *Management Science* **59**(1) 157-171.

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

Andritsos, D. A., C. S. Tang. 2014. Introducing competition in healthcare services: The role of private care and increased patient mobility. *European Journal of Operational Research* 234(3) 898-909.

Andritsos, D. A., C. S. Tang. 2015. Incentive programs for reducing readmissions when patient care is co-produced. *Working Paper*, UCLA Anderson School.

Ata, B., B. L. Killaly, T. L. Olsen, R. P. Parker. 2013. On hospice operations under medicare reimbursement policies. *Management Science* **59**(5) 1027-1044.

Augenstein, S., S. K. Livio. 2012. Moody's predicts more N.J. hospitals will merge, seek buyers. Available at http://www.nj.com/news/index.ssf/2012/02/moodys_predicts_more_nj_hospit.html (accessed date February 3, 2012).

Australian hospital statistics report. 2015. Elective surgery waiting times 2014-15: Australian hospital statistics. Available at http://www.aihw.gov.au/publication-detail/?id =60129553174 (accessed date October 15, 2015).

Australian Primary Health Care Advisory Group Report. 2015. Better outcomes for people with chronic and complex health conditions. Available at http://www.health.gov.au/internet/main/publishing.nsf/Content/primary-phcag-report.

Bavafa, H., S. Savin, C. Terwiesch. 2013. Managing office readmission intervals and patient panel sizes in primary care. Working Paper.

Barua, B., F. Fathers. 2014. Waiting your turn: wait times for health care in Canada 2014 report. Studies in Health Policy. Vancouver: Fraser Institute.

Blomqvist, A., C. Busby. 2013. Paying hospital-based doctors: fee for whose service? Commentary 392. Toronto: C.D. Howe Institute.

Calsyn, M., E. O. Lee. 2012. Alternatives to Fee-for-Service payments in health care. *Center for American Progress.*

Campbell, D. 2014. NHS patients waiting longer for routine operations under coalition. Available at http://www.theguardian.com/society/2014/jul/04/nhs-patients-waiting-longer-for-routine-operations-under-coalition (accessed date July 4, 2014).

Carey, K. 2015. Measuring the hospital length of stay/readmission cost trade-off under a bundled payment mechanism. *Health Economics* **24** 790-802.

Chan, C. W., G. B. Yom-Tov, G. Escobar. 2014. When to use speedup: an examination of service systems with returns. *Operations Research* **62**(2) 462-482.

Dai, T., M. Akan, S. Tayur. 2017. Imaging room and beyond: the underlying economics behind physicians' test-ordering behavior in outpatient services. *Manufacturing & Service Operations Management* **19**(1) 99-113.

Davis, K. 2007. Paying for care episodes and care coordination. *The New England Journal of Medicine* **356**(11) 1166-1168.

DiChiara, J. 2015. Rural hospitals address medicare reimbursement cut concerns. Available at http://revcycleintelligence.com/news/rural-hospitals-address-medicare-reimbursement-cut-concerns (accessed date May 12, 2015).

Dimakou, S. 2013. Waiting time distributions and national targets for elective surgery in UK: theoretical modelling and duration analysis. Doctoral thesis, Department of Economics, City University London, UK.

Donnelly, L. 2013. NHS is about to run out of cash top official warns. Available at http://www.telegraph.co.uk/news/health/news/10162848/NHS-is-about-to-run-out-of-cash-top-official-warns.html (accessed date July 5, 2013).

de Vericourt, F., Y. Zhou. 2005. Managing response time in a call-routing problem with service failure. *Operations Research* **53**(6) 968-981.

Fenter, T., S. Lewis. 2008. Pay-for-performance initiatives. *Journal of Managed Care Pharmacy* **14**(6)(suppl S-c) s12-s15.

Fethke, C. C., I. M. Smith, N. Johnson. 1986. "Risk" factors affecting readmission of the elderly into the health care system. *Medical Care* **24**(5) 429-437.

Fleck, F. 2002. Governments need to be more aware of human rights implications of health care. *British Medical Journal* **325**(7358) 238.

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

General Practice Forward View. 2016. Five year forward view for general practice published by NHS England. Available at http://nhsconfed.org/resources/2016/05/five-year-forward-view-for-general-practice-published-by-nhs-england.

Guo, P., R. Lindsey, Q. Qian. 2016. Efficiency of subsidy schemes in reducing waiting times for public health-care services. Working paper, Sauder School of Business, University of British Columbia, Vancouver.

Guo, P., C. S. Tang, Y. Wang, M. Zhao. 2017. The impact of reimbursement policy on social welfare, revisit rate and waiting time in a public healthcare system: Fee-for-Service vs. Bundled Payment. Working Paper, the Hong Kong Polytechnic University. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3058734.

Gupta, D., M. Mehrotra. 2015. Bundled payments for healthcare services: proposer selection and information sharing. *Operations Research* **63**(4) 772-788.

Hasija, S., E. Pinker, R. A. Shumsky. 2009. Work expands to fill the time available: Capacity estimation and staffing under parkinson's law. *Manufacturing & Service Operations Management* **12**(1) 1-18.

Hassin, R., M. Haviv. 2003. To queue or not to queue: equilibrium behavior in queueing systems. Kluwer Academic Publishers, Norwell, MA.

Hopp, W. J., S. M. R. Iravani, G. Y. Yuen. 2007. Operations systems with discretionary task completion. *Management Science* **53**(1) 61-77.

Hing, E., C. J. Hsiao. 2014. State variability in supply of office-based primary care providers: United States, 2012. *NCHS Data Brief* **151**.

Hurst, J., L. Siciliani. 2003. Tackling excessive waiting times for elective surgery: a comparison of policies in twelve OECD countries. OECD Health Working Papers, No 6, Paris.

Japsen, B. 2015. Medicare bundled payment gains momentum with hospitals, nursing homes. *Forbes*. Published on August 21, 2015.

Jiang, H., Z. Pang, S. Savin. 2012. Performance-based contracts for outpatient medical services. *Manufacturing & Service Operations Management* **14**(4) 654-669.

Kim, S., I. Horowitz, K. K. Young, T. A. Buckley. 1999. Analysis of capacity management of the intensive care unit in a hospital. *European Journal of Operational Research* **115**(1) 36-46.

Konrad, T. R., C. L. Link, R. J. Shackelton et al. 2010. It's about time: physicians' perceptions of time constraints in primary care medical practice in three national healthcare systems. *Med Care* **48**(2) 95-100.

Kostami, V., S. Rajagopalan. 2014. Speed quality tradeoff in a dynamic model. *Manufacturing & Service Operations Management* **16**(1) 104-118.

Lee, DK. K., S. A. Zenios. 2012. An evidence-based incentive system for Medicare's End-Stage Renal Disease program. *Management Science* **58**(6) 1092-1105.

Levesque, J. F., M. F. Harris, G. Russell. 2013. Patient-centred access to health care: conceptualising access at the interface of health systems and populations. *International Journal for Equity in Health* **12**(18) 16-28.

Li, X., P. Guo, Z. Lian. 2016. Speed-quality competition in customer-intensive services with boundedly rational customer. *Production and Operations Management* **25**(11) 1885-1901.

Ministry of Health and Long-Term Care. 2015. Ontario funds bundled care teams to improve patient experience. Available at https://news.ontario.ca/mohltc/en/2015/09/ontario-funds-bundled-care-teams-to-improve-patient-experience.html (accessed date September 2, 2015).

Morrow-Howell, N., E. K. Proctor. 1993. The use of logistic regression in social work research. *Journal of Social Service Research* **16**(1-2) 87-104.

Mot, E. S. 2002. Paying the medical specialist: the eternal puzzle: experiments in the Netherlands. PhD Thesis page 176, Amsterdam.

Official Website of the South Australian Government. Elective surgery services. Available at http://www.sahealth.sa.gov.au/wps/wcm/connect/Public+Content/SA+Health+Internet/Health+services/Elective+surgery+services/.

Official Website of Hong Kong Hospital Authority. 2016. Waiting time for cataract surgery. Available at http://www.ha.org.hk/visitor/ha_visitor_text_index.asp?Parent_ID=214172&Content_ID=214184.

Paç, M. F., S. Veeraraghavan. 2010. Strategic diagnosis and pricing in expert services. Working paper, Wharton School, University of Pennsylvania, Philadelphia.

Palacios, M., B. Barua, F. Ren. 2015. The price of public health care insurance. *Fraser Research Bulletin*. Published on August 2015.

Plambeck, E. L., S. A. Zenios. 2000. Performance-based incentives in a dynamic principal-agent model. *Manufacturing & Service Operations Management* **2**(3) 240-263.

Rabin, R. 2014. 15-minute Doctor Visits Take a Toll on a Patient-Physician Relationship. *PBS Newshour*. April 21.

Ross, S. 2007. Introduction to probability models. Academic Press, USA.

So, K. C., C. S. Tang. 2000. Modeling the impact of an outcome-oriented reimbursement

policy on clinic, patients, and pharmaceutical firms. *Management Science* **46**(7) 875-892.

Street, A., J. O'Reilly, P. Ward, A. Mason. 2011. DRG-based hospital payment and efficiency: Theory, evidence, and challenges. In *Diagnosis-Related Groups in Europe - Moving Towards Transparency, Efficiency and Quality in Hospitals*, Busse R, Geissler, A, Quentin W, Wiley M (eds). Open University Press: Maidenhead; 93-114.

Tsai, T. C., K. E. Joynt, R. C. Wild, E. J. Orav, A. K. Jha. 2015. Medicare's Bundled Payment initiative: most hospitals are focused on a few high-volume conditions. *Health Affairs* **34**(3) 371-380.

Tong, C., S. Rajagopalan. 2014. Pricing and operational performance in discretionary services. *Production and Operations Management* **23**(4) 689-703.

Varkevisser, M., S. A. van der Geest, F. T. Schut. 2012. Do patients choose hospitals with high quality ratings? Empirical evidence from the market for angioplasty in the Netherlands. Journal of Health Economics **31**(2) 371-378.

van der Linden, W., A. Warg, P. Nordin. 2011. National register study of operating time and outcome in hernia repair. *Arch Surg.* **146**(10) 1198-1203.

Wang, X., L. G. Debo, A. Scheller-Wolf, S. F. Smith. 2010. Design and analysis of diagnostic service centers. *Management Science* **56**(11) 1873-1890.

Xu, Y., A. Scheller-Wolf, K. Sycara. 2015. The benefit of introducing variability in quality based service domains. *Operations Research* **63**(1) 233-246.

Yaesoubi, R., S. D. Roberts. 2011. Payment contracts in a preventive health care system: A perspective from Operations Management. *Journal of Health Economics* **30**(6) 1188-1196.

Yom-Tov, G. B., A. Mandelbaum. 2014. Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management* **16**(2) 283-299.

Online Appendix

"The Impact of Reimbursement Policy on Social Welfare, Revisit Rate and Waiting Time in a Public Healthcare System: Fee-for-Service vs. Bundled Payment"

P. Guo, C. Tang, Y. Wang, M. Zhao

Proof of Lemma 1. Taking the first-order condition (FOC) of $o(\mu)$ over μ yields

$$\frac{do(\mu)}{d\mu} = 1 - \delta(\mu) - \mu\delta^{'}(\mu) = 0,$$

which can be rewritten as $(1-\delta(\mu))(1-\mu g(\mu)) = 0$. As $g(\mu)$ is increasing in μ, $do(\mu)/d\mu$ crosses zero only once from above. Therefore, $o(\mu)$ is unimodal in μ, and the optimal service rate μ^o solves $\mu^o g(\mu^o) = 1$.

Next, we prove that when $\mu \leq \mu^o$, $o(\mu)$ is concave in μ. As $g(\mu)$ is increasing in μ,

$$\frac{dg(\mu)}{d\mu} = \frac{\delta^{''}(\mu)(1-\delta(\mu)) + (\delta^{'}(\mu))^2}{(1-\delta(\mu))^2} > 0, \tag{23}$$

which yields $\delta^{''}(\mu)(1-\delta(\mu)) + (\delta^{'}(\mu))^2 > 0$. Then we can show that

$$\begin{aligned}
\frac{d^2o(\mu)}{d\mu^2} &= -2\delta^{'}(\mu) - \mu\delta^{''}(\mu) \\
&= -2\delta^{'}(\mu) + \frac{\mu(\delta^{'}(\mu))^2}{1-\delta(\mu)} - \frac{\mu\delta^{''}(\mu)(1-\delta(\mu)) + \mu(\delta^{'}(\mu))^2}{1-\delta(\mu)} \\
&= -\delta^{'}(\mu) - \frac{\delta^{'}(\mu)}{1-\delta(\mu)}\frac{do(\mu)}{d\mu} - \frac{\mu\delta^{''}(\mu)(1-\delta(\mu)) + \mu(\delta^{'}(\mu))^2}{1-\delta(\mu)}.
\end{aligned}$$

As $o(\mu)$ is increasing in μ when $\mu \leq \mu^o$, the above equation is negative for all $\mu \leq \mu^o$. Therefore, $o(\mu)$ is concave in μ for $\mu \leq \mu^o$. □

Proof of Corollary 1. Denote $\mu_i = \max\{\tilde{\lambda}(\mu)\}$ and $\mu_y = \max\{\tilde{\lambda}_e(\mu)\}$. By noting that $o(\mu) = \mu(1-\delta(\mu))$ and $o^{'}(\mu) = 1-\delta(\mu)-\mu\delta^{'}(\mu)$, we have $o(\mu)-\mu o^{'}(\mu) = \mu^2\delta^{'}(\mu)$. Therefore, taking the derivative of $\tilde{\lambda}(\mu)$ with respect to μ yields

$$\frac{d\tilde{\lambda}(\mu)}{d\mu} = o^{'}(\mu) - \frac{\theta o^{'}(\mu)}{Ro(\mu)-\mu t} + \frac{\theta o(\mu)(Ro^{'}(\mu)-t)}{[Ro(\mu)-\mu t]^2} = o^{'}(\mu) - \frac{\theta\delta^{'}(\mu)t}{[R(1-\delta(\mu))-t]^2}. \tag{24}$$

From Lemma 1, we can show that $d\tilde{\lambda}(\mu)/d\mu < 0$ for $\mu \geq \mu^o$. Thus, $\mu_i < \mu^o$. Using Lemma

1 again, $o^{''}(\mu_i) < 0$. Recall from (23) that $\delta^{''}(\mu)(1-\delta(\mu)) + (\delta^{'}(\mu))^2 > 0$. Therefore,

$$\begin{aligned}
\left.\frac{d^2\tilde{\lambda}(\mu)}{d\mu^2}\right|_{\mu=\mu_i} &= o^{''}(\mu_i) - \frac{\theta t\delta^{''}(\mu_i)}{[R(1-\delta(\mu_i))-t]^2} - \frac{2R\theta(\delta^{'}(\mu_i))^2 t}{[R(1-\delta(\mu_i))-t]^3} \\
&= o^{''}(\mu_i) + \frac{\theta t(\delta^{'}(\mu_i))^2}{(1-\delta(\mu_i))[R(1-\delta(\mu_i))-t]^2} - \frac{2R\theta(\delta^{'}(\mu_i))^2 t}{[R(1-\delta(\mu_i))-t]^3} \\
&\quad -\frac{\theta t[\delta^{''}(\mu_i)(1-\delta(\mu_i)) + (\delta^{'}(\mu_i))^2]}{(1-\delta(\mu_i))[R(1-\delta(\mu_i))-t]^2} \\
&= o^{''}(\mu_i) - \frac{\theta t[\delta^{''}(\mu_i)(1-\delta(\mu_i)) + (\delta^{'}(\mu_i))^2]}{(1-\delta(\mu_i))[R(1-\delta(\mu_i))-t]^2} - \frac{\theta t(\delta^{'}(\mu_i))^2(R(1-\delta(\mu_i))+t)}{(1-\delta(\mu_i))[R(1-\delta(\mu_i))-t]^3} \quad (25) \\
&< 0,
\end{aligned}$$

which shows that $\tilde{\lambda}(\mu)$ is unimodal in μ. Taking the derivative of $\tilde{\lambda}_e(\mu)$ with respect to μ and using (23), we have

$$\begin{aligned}
\frac{d\tilde{\lambda}_e(\mu)}{d\mu} &= 1 - \frac{\theta R\delta^{'}(\mu)}{[R(1-\delta(\mu))-t]^2}; \quad (26) \\
\frac{d^2\tilde{\lambda}_e(\mu)}{d\mu^2} &= -\frac{\theta R\delta^{''}(\mu)}{[R(1-\delta(\mu))-t]^2} - \frac{2\theta R^2(\delta^{'}(\mu))^2}{[R(1-\delta(\mu))-t]^3} \\
&= \frac{-R\theta}{(1-\delta(\mu))}\left[\frac{(\delta^{'}(\mu))^2(R(1-\delta(\mu))+t)}{[R(1-\delta(\mu))-t]^3} + \frac{\delta^{''}(\mu)(1-\delta(\mu)) + (\delta^{'}(\mu))^2}{[R(1-\delta(\mu))-t]^2}\right] \\
&< 0. \quad (27)
\end{aligned}$$

Thus, $\tilde{\lambda}_e(\mu)$ is concave in μ. We next show that the mode of $\tilde{\lambda}(\mu)$ is smaller than the mode of $\tilde{\lambda}_e(\mu)$, that is, $\mu_i < \mu_y$. By noting that $o^{'}(\mu) = 1-\delta(\mu) - \mu\delta^{'}(\mu)$, (26) can be rewritten as

$$\frac{d\tilde{\lambda}_e(\mu)}{d\mu} = 1 - \frac{\theta}{\mu[R(1-\delta(\mu))-t]} + \frac{\theta(Ro^{'}(\mu)-t)}{\mu[R(1-\delta(\mu))-t]^2} = \frac{\tilde{\lambda}_e(\mu)}{\mu} + \frac{\theta(Ro^{'}(\mu)-t)}{\mu[R(1-\delta(\mu))-t]^2}.$$

Obviously, the maximum effective visit rate should be positive, that is, $\tilde{\lambda}_e(\mu_y) > 0$. Because μ_y should satisfy the FOC of $\tilde{\lambda}_e(\mu)$, namely, $d\tilde{\lambda}_e(\mu_y)/d\mu = 0$, we have $Ro^{'}(\mu_y) - t < 0$. Using (26) and substituting $d\tilde{\lambda}_e(\mu_y)/d\mu = 0$ into (24), we have

$$\left.\frac{d\lambda(\mu)}{d\mu}\right|_{\mu=\mu_y} = \frac{Ro^{'}(\mu_y)-t}{R} < 0, \quad (28)$$

which implies that $\mu_y > \mu_i$. □

Proof of Proposition 1. With a slight abuse of notation, we interchangeably use $\tilde{\mu}_s(r_s)$ and $\tilde{\mu}_s$. From (10), the FOC of $\Pi_f(\mu)$ can be written as

$$\frac{d\Pi_f(\mu)}{d\mu} = \frac{c\tilde{\lambda}_e(\mu)}{\mu^2} + \left(r_f - \frac{c}{\mu}\right)\frac{d\tilde{\lambda}_e(\mu)}{d\mu} = 0, \quad (29)$$

which can be rewritten as (12). Furthermore, the optimal service rate $\tilde{\mu}_f(r_f)$ should satisfy $\frac{d\tilde{\lambda}_e(\mu)}{d\mu}|_{\mu=\tilde{\mu}_f} < 0$. By recalling from Corollary 1 that $\tilde{\lambda}_e(\mu)$ is concave in μ, $\tilde{\mu}_f(r_f)$ is larger than the mode of $\tilde{\lambda}_e(\mu)$. Furthermore,

$$\left.\frac{d^2\Pi_f(\mu)}{d\mu^2}\right|_{\mu=\tilde{\mu}_f} = \left.\frac{2c}{\tilde{\mu}_f^2}\frac{d\tilde{\lambda}_e(\mu)}{d\mu}\right|_{\mu=\tilde{\mu}_f} - \frac{2c\tilde{\lambda}_e(\tilde{\mu}_f)}{\tilde{\mu}_f^3} + \left(r_f - \frac{c}{\tilde{\mu}_f}\right)\left.\frac{d^2\tilde{\lambda}_e(\mu)}{d\mu^2}\right|_{\mu=\tilde{\mu}_f} < 0,$$

which shows that $\Pi_f(\mu)$ is unimodal in μ, and therefore $\tilde{\mu}_f$ maximizes $\Pi_f(\mu)$.

From (11), taking the derivative of $\Pi_b(\mu)$ with respect to μ yields

$$\frac{d\Pi_b(\mu)}{d\mu} = \frac{co^{'}(\mu)\tilde{\lambda}(\mu)}{(o(\mu))^2} + \left(r_b - \frac{c}{o(\mu)}\right)\frac{d\tilde{\lambda}(\mu)}{d\mu}. \tag{30}$$

From (24) and Lemma 1, we can know that $o^{'}(\mu) < 0$ and $d\tilde{\lambda}(\mu)/d\mu < 0$ for $\mu > \mu^o$. To ensure that $\Pi_b(\mu) > 0$, $r_b > c/o(\mu)$ is required. Hence, if $\tilde{\mu}_b \geq \mu^o$, $d\Pi_b(\mu)/d\mu < 0$. Therefore, the optimal service rate selected by the HCP satisfies $\tilde{\mu}_b < \mu^o$. According to Lemma 1, we then have $o^{'}(\tilde{\mu}_b) > 0$ and $o^{''}(\tilde{\mu}_b) < 0$. Then from (25) and (30), we can get

$$\frac{d^2\tilde{\lambda}(\tilde{\mu}_b)}{d\mu^2} < 0; \frac{d\tilde{\lambda}(\tilde{\mu}_b)}{d\mu} < 0.$$

As $\tilde{\lambda}(\mu)$ is unimodal in μ, $\tilde{\mu}_b$ is larger than the mode of $\tilde{\lambda}(\mu)$. From (24), we have $\frac{d\tilde{\lambda}(\mu^o)}{d\mu} < 0$ as $o^{'}(\mu^o) = 0$. Because $\tilde{\mu}_b < \mu^o$ and $\tilde{\lambda}(\mu)$ is unimodal in μ, we have $\tilde{\lambda}(\mu^o) < \tilde{\lambda}(\tilde{\mu}_b)$. Furthermore,

$$\left.\frac{d^2\Pi_b(\mu)}{d\mu^2}\right|_{\mu=\tilde{\mu}_b} = \frac{c(o^{''}(\tilde{\mu}_b)o(\tilde{\mu}_b) - 2(o^{'}(\tilde{\mu}_b))^2)\tilde{\lambda}(\tilde{\mu}_b)}{(o(\tilde{\mu}_b))^3} + \frac{2co^{'}(\tilde{\mu}_b)}{(o(\tilde{\mu}_b))^2}\frac{d\tilde{\lambda}(\tilde{\mu}_b)}{d\mu} + \left(r_b - \frac{c}{o(\tilde{\mu}_b)}\right)\frac{d^2\tilde{\lambda}(\tilde{\mu}_b)}{d\mu^2} < 0.$$

Therefore, $\Pi_b(\mu)$ is unimodal in μ and the optimal service rate $\tilde{\mu}_b$ satisfies $d\Pi_b(\tilde{\mu}_b)/d\mu = 0$, which can be rewritten as (13). □

Proof of Corollary 2. With a slight abuse of notation, we interchangeably use $\tilde{\mu}_s(r_s)$ and $\tilde{\mu}_s$. Differentiating (29) with respect to r_f, we have $\frac{\partial^2\Pi_f(\mu)}{\partial\mu\partial r_f} = \frac{d\tilde{\lambda}_e(\mu)}{d\mu}$. Recall from the proof of Proposition 1 that $\frac{d\tilde{\lambda}_e(\mu)}{d\mu}|_{\mu=\tilde{\mu}_f} < 0$. According to the implicit function theorem, $\frac{d\tilde{\mu}_f}{dr_f} = -\left.\frac{\frac{\partial^2\Pi_f(\mu)}{\partial\mu\partial r_f}}{\frac{\partial^2\Pi_f(\mu)}{\partial\mu^2}}\right|_{\mu=\tilde{\mu}_f} < 0$. Therefore, $\tilde{\mu}_f$ is decreasing in r_f. As $\delta(\mu)$ is increasing in μ, $\delta(r_f)$ is also decreasing in r_f. Next, substituting $\tilde{\mu}_f$ into (9) and taking the derivative of $\tilde{\lambda}_e(r_f)$ over r_f yields

$$\frac{d\tilde{\lambda}_e(r_f)}{dr_f} = \frac{d\tilde{\lambda}_e(\tilde{\mu}_f)}{d\mu}\frac{d\tilde{\mu}_f}{dr_f} > 0.$$

By using (3),

$$\frac{d\tilde{\lambda}(r_f)}{dr_f} = -\delta^{'}(\tilde{\mu}_f)\frac{d\tilde{\mu}_f}{dr_f}\tilde{\lambda}_e(\tilde{\mu}_f) + (1-\delta(\tilde{\mu}_f))\frac{d\tilde{\lambda}_e(r_f)}{dr_f} > 0.$$

Differentiating (30) with respect to r_b, we have $\frac{\partial^2 \Pi_b(\mu)}{\partial \mu \partial r_b}\big|_{\mu=\tilde{\mu}_b} = \frac{d\tilde{\lambda}(\tilde{\mu}_b)}{d\mu}$. Recall from the proof of Proposition 1 that $\frac{d\tilde{\lambda}(\tilde{\mu}_b)}{d\mu} < 0$ and $\frac{d^2\Pi_b(\tilde{\mu}_b)}{d\mu^2} < 0$. Using the implicit function theorem again, we have $\frac{d\tilde{\mu}_b}{dr_b} = -\frac{\frac{\partial^2 \Pi_b(\mu)}{\partial \mu \partial r_b}}{\frac{\partial^2 \Pi_b(\mu)}{\partial \mu^2}}\Big|_{\mu=\tilde{\mu}_b} < 0$. Therefore, $\tilde{\mu}_b$ is decreasing in r_b. As $\delta(\mu)$ is increasing in μ, $\delta(r_b)$ is also decreasing in r_b. Next, substituting $\tilde{\mu}_b$ into (8) and taking the derivative of $\tilde{\lambda}(r_b)$ over r_b yields

$$\frac{d\tilde{\lambda}(r_b)}{dr_b} = \frac{d\tilde{\mu}_b}{dr_b}\frac{d\tilde{\lambda}(\tilde{\mu}_b)}{d\mu} > 0.$$

Finally, plugging $\tilde{\lambda}_e(\tilde{\mu}_s)$ into (5) and (6), we get $W(r_s) = \frac{R(1-\delta(\tilde{\mu}_s))}{\theta} - \frac{t}{\theta}$ and $T(r_s) = \frac{R}{\theta} - \frac{t}{\theta(1-\delta(\tilde{\mu}_s))}$. Differentiating $W(r_s)$ and $T(r_s)$ with respect to r_s yields

$$\frac{dW(r_s)}{dr_s} = -\frac{R\delta^{'}(\tilde{\mu}_s)}{\theta}\frac{d\tilde{\mu}_s}{dr_s} > 0;\ \frac{dT(r_s)}{dr_s} = -\frac{t\delta^{'}(\tilde{\mu}_s)}{\theta(1-\delta(\tilde{\mu}_s))^2}\frac{d\tilde{\mu}_s}{dr_s} > 0.$$

□

Proof of Proposition 2. According to Corollary 2, this proposition is immediate. □

Proof of Corollary 3. According to Proposition 2, $\tilde{r}_f\tilde{\lambda}_e(\tilde{r}_f) = B$ and $\tilde{r}_b\tilde{\lambda}(\tilde{r}_b) = B$. Recall from the proof of Corollary (2) that $\tilde{\lambda}_e(r_f)$ is increasing in r_f, and $\tilde{\lambda}(r_b)$ is increasing in r_b. Therefore, as B increases, $\tilde{r}_s$ should increase. Using Corollary (2) again, we can easily show that the total waiting time $T(\tilde{r}_s)$, the waiting time per visit $W(\tilde{r}_s)$, and the initial visit rate $\tilde{\lambda}(\tilde{r}_s)$ are increasing in B, while $\tilde{\mu}_s(\tilde{r}_s)$ and $\delta(\tilde{r}_s)$ are decreasing in B. □

Proof of Proposition 3. From (17), maximizing $S(\mu)$ is equivalent to maximizing $\tilde{\lambda}(\mu)$. According to Corollary 1, $S(\mu)$ is unimodal in μ. Let $\mu_i = argmax\{\tilde{\lambda}(\mu)\}$. When $B \geq c \cdot \tilde{\lambda}_e(\mu_i)/\mu_i$, μ_i is feasible and so $\mu^* = \mu_i = argmax\{\tilde{\lambda}(\mu)\}$. Otherwise, if $B < c \cdot \tilde{\lambda}_e(\mu_i)/\mu_i$, μ_i is infeasible and the budget constraint (18) is binding. We next show that $\tilde{\lambda}_e(\mu)/\mu$ is unimodal in μ and so there exists two service rates that make the the budget constraint (18) binding. From (9), we have

$$\frac{\tilde{\lambda}_e(\mu)}{\mu} = 1 - \frac{\theta}{R \cdot o(\mu) - t\mu}.$$

Taking the derivative of $\tilde{\lambda}_e(\mu)/\mu$ with respect to μ yields

$$\frac{d}{d\mu}\left(\frac{\tilde{\lambda}_e(\mu)}{\mu}\right) = \frac{\theta(R \cdot o^{'}(\mu) - t)}{(R \cdot o(\mu) - t\mu)^2}.$$

If $\frac{d}{d\mu}\left(\frac{\tilde{\lambda}_e(\mu)}{\mu}\right)$ vanishes at μ_e, then $o^{'}(\mu_e) > 0$. According to Lemma 1, $o^{''}(\mu_e) < 0$. Therefore,

$$\frac{d^2}{d\mu^2}\left(\frac{\tilde{\lambda}_e(\mu)}{\mu}\right)\Big|_{\mu=\mu_e} = \frac{\theta(R \cdot o^{''}(\mu_e) - t)}{(R \cdot o(\mu_e) - t\mu_e)^2} < 0,$$

which implies that $\tilde{\lambda}_e(\mu)/\mu$ is unimodal in μ and there are two solutions of $c \cdot \tilde{\lambda}_e(\mu)/\mu = B$. Let μ_1 and μ_2 denote the two solutions, respectively. Without loss of generality, we assume that $\mu_2 > \mu_1$. We next show that $o(\mu_2) > o(\mu_1)$. Suppose this is not true such that $o(\mu_2) \leq o(\mu_1)$. Then, $R \cdot o(\mu_2) - t\mu_2 < R \cdot o(\mu_1) - t\mu_1$ and so

$$\frac{\tilde{\lambda}_e(\mu_2)}{\mu_2} = 1 - \frac{\theta}{R \cdot o(\mu_2) - t\mu_2} < 1 - \frac{\theta}{R \cdot o(\mu_1) - t\mu_1} = \frac{\tilde{\lambda}_e(\mu_1)}{\mu_2},$$

which contradicts the definition of μ_1 and μ_2. Therefore, $o(\mu_2) > o(\mu_1)$. From (3), we have $\frac{\tilde{\lambda}_e(\mu)}{\mu} = \frac{\tilde{\lambda}(\mu)}{o(\mu)}$. Because $\frac{c \cdot \tilde{\lambda}_e(\mu_1)}{\mu_1} = \frac{c \cdot \tilde{\lambda}_e(\mu_2)}{\mu_2} = B$ and $o(\mu_2) > o(\mu_1)$, $\tilde{\lambda}(\mu_1) < \tilde{\lambda}(\mu_2)$. As maximizing $S(\mu)$ is equivalent to maximizing $\tilde{\lambda}(\mu)$, $\mu^* = \mu_2$. In other words, when $B < c \cdot \tilde{\lambda}_e(\mu_i)/\mu_i$, the optimal solution μ^* is the larger service rate that satisfies $c \cdot \tilde{\lambda}_e(\mu)/\mu = B$. Finally, to facilitate our proof of Proposition 4, we next show that when $B < c \cdot \tilde{\lambda}_e(\mu_i)/\mu_i$, $\mu^* > \mu_i > \mu_e$. Because μ_i is infeasible in this case, $\mu_i < \mu^*$. Furthermore, because $o^{'}(\mu_e) > 0$ and

$$\frac{d}{d\mu}\left(\frac{\tilde{\lambda}_e(\mu)}{\mu}\right)\Big|_{\mu=\mu_e} = \frac{1}{o(\mu_e)}\frac{d\tilde{\lambda}(\mu)}{d\mu}\Big|_{\mu=\mu_e} - \frac{o^{'}(\mu_e)\tilde{\lambda}(\mu_e)}{o^2(\mu_e)} = 0,$$

we can show that $\frac{d\tilde{\lambda}(\mu)}{d\mu}\big|_{\mu=\mu_e} > 0$. Because $\tilde{\lambda}(\mu)$ is unimodal in μ and $\mu_i = argmax\{\tilde{\lambda}(\mu)\}$, $\mu_i > \mu_e$. □

Proof of Proposition 4. We first compare the BP scheme with the benchmark case. According to Proposition 3, when $B \geq c \cdot \tilde{\lambda}_e(\mu_i)/\mu_i$, $\mu^* = \mu_i = argmax\{\tilde{\lambda}(\mu)\}$, while when $B < c \cdot \tilde{\lambda}_e(\mu_i)/\mu_i$, μ^* is the larger service rate that satisfies $c \cdot \tilde{\lambda}_e(\mu)/\mu = B$. From the proof of Proposition 1, we can know that $\frac{d\tilde{\lambda}(\tilde{\mu}_b(\tilde{r}_b))}{d\mu} < 0$. Therefore, from Corollary 1, when $B \geq c \cdot \tilde{\lambda}_e(\mu_i)/\mu_i$, $\mu^* = \mu_i < \tilde{\mu}_b(\tilde{r}_b)$, $\tilde{\lambda}(\mu^*) = \tilde{\lambda}(\mu_i) > \tilde{\lambda}(\tilde{\mu}_b(\tilde{r}_b)) = \tilde{\lambda}(\tilde{r}_b)$, and $S(\mu^*) = S(\mu_i) > S(\tilde{\lambda}(\tilde{\mu}_b(\tilde{r}_b))) = S(\tilde{r}_b)$. We next consider the scenario $B < c \cdot \tilde{\lambda}_e(\mu_i)/\mu_i$. Under the BP scheme, the HCP's profit in equilibrium must be positive, which implies that $\tilde{r}_b\tilde{\lambda}(\tilde{\mu}_b(\tilde{r}_b)) > \frac{c \cdot \tilde{\lambda}_e(\tilde{\mu}_b(\tilde{r}_b))}{\tilde{\mu}_b(\tilde{r}_b)}$. Since $\tilde{r}_b\tilde{\lambda}(\tilde{\mu}_b(\tilde{r}_b)) = B$ (according to Proposition 2), $\frac{c \cdot \tilde{\lambda}_e(\tilde{\mu}_b(\tilde{r}_b))}{\tilde{\mu}_b(\tilde{r}_b)} < B$. Thus, when $B < c \cdot \tilde{\lambda}_e(\mu_i)/\mu_i$, $\frac{c \cdot \tilde{\lambda}_e(\tilde{\mu}_b(\tilde{r}_b))}{\tilde{\mu}_b(\tilde{r}_b)} < \frac{c \cdot \tilde{\lambda}_e(\mu^*)}{\mu^*}$. Recall from the proof of Proposition 3 that $\frac{\tilde{\lambda}_e(\mu)}{\mu}$ is unimodal in μ and when $B < c \cdot \tilde{\lambda}_e(\mu_i)/\mu_i$, $\mu^* > \mu_i > \mu_e$. Because $\tilde{\mu}_b(\tilde{r}_b) > \mu_i$ (according to Proposition 1) and $\tilde{\lambda}(\mu)$ is unimodal in μ, we have $\tilde{\mu}_b(\tilde{r}_b) > \mu^*$ and $\tilde{\lambda}(\mu^*) > \tilde{\lambda}(\tilde{\mu}_b(\tilde{r}_b)) = \tilde{\lambda}(\tilde{r}_b)$. Under the partial coverage, the funder's objective is equivalent to maximizing the initial visit rate. Therefore, $S(\mu^*) > S(\tilde{\mu}_b(\tilde{r}_b)) = S(\tilde{r}_b)$. Because $\delta(\mu)$ is increasing in μ and $\mu^* < \tilde{\mu}_b(\tilde{r}_b)$, $\delta(\mu^*) < \delta(\tilde{\mu}_b(\tilde{r}_b)) = \delta(\tilde{r}_b)$.

We next compare the FFS scheme with the BP scheme. To this end, we first show $\tilde{\lambda}(\tilde{r}_f) < \tilde{\lambda}(\tilde{r}_b)$ by contradiction. Suppose this is not true so that $\tilde{\lambda}(\tilde{r}_f) \geq \tilde{\lambda}(\tilde{r}_b)$. Recall from Corollary 2 that $\tilde{\lambda}(r_f)$ is increasing in r_f. As $\tilde{\mu}_f(r_f)$ is decreasing in r_f and

$$\frac{d\tilde{\lambda}(r_f)}{dr_f} = \frac{d\tilde{\lambda}(\tilde{\mu}_f(r_f))}{d\mu}\frac{d\tilde{\mu}_f(r_f)}{dr_f},$$

we can obtain

$$\frac{d\tilde{\lambda}(\tilde{\mu}_f(r_f))}{d\mu} < 0.$$

From the proof of Proposition 1 we have $\frac{d\tilde{\lambda}(\tilde{\mu}_b(\tilde{r}_b))}{d\mu} < 0$. Because $\tilde{\lambda}(\mu)$ is unimodal in μ and $\tilde{\lambda}(\tilde{r}_f) = \tilde{\lambda}(\tilde{\mu}_f(\tilde{r}_f)) \geq \tilde{\lambda}(\tilde{\mu}_b(\tilde{r}_b)) = \tilde{\lambda}(\tilde{r}_b)$, this implies that $\tilde{\mu}_f(\tilde{r}_f) \leq \tilde{\mu}_b(\tilde{r}_b)$. Furthermore, according to Proposition 2, under the partial coverage, the budget constraints associated with both the FFS and BP schemes are binding, that is, $\tilde{r}_f\tilde{\lambda}_e(\tilde{r}_f) = \tilde{r}_b\tilde{\lambda}(\tilde{r}_b) = B$. Recall that $\tilde{\lambda}_e(\tilde{r}_f) = \tilde{\lambda}(\tilde{r}_f)/(1-\delta(\tilde{\mu}_f(\tilde{r}_f)))$. Then we have

$$\frac{\tilde{r}_f}{\tilde{r}_b} = \frac{\tilde{\lambda}(\tilde{r}_b)}{\tilde{\lambda}_e(\tilde{r}_f)} = \frac{\tilde{\lambda}(\tilde{r}_b)}{\tilde{\lambda}(\tilde{r}_f)}(1-\delta(\tilde{\mu}_f(\tilde{r}_f)) \leq 1-\delta(\tilde{\mu}_f(\tilde{r}_f))$$

because $\tilde{\lambda}(\tilde{r}_f) \geq \tilde{\lambda}(\tilde{r}_b)$. Taking a close look at (10) and (11), when $r_b = \tilde{r}_b$, we can utilize (3) to rewrite the profit function of the HCP under the BP scheme as

$$\Pi_b(\mu) = \left(\tilde{r}_b - \frac{\tilde{r}_f}{1-\delta(\mu)}\right)\tilde{\lambda}(\mu) + \Pi_f(\mu). \tag{31}$$

Note that $\tilde{\mu}_f(\tilde{r}_f)$ maximizes $\Pi_f(\mu)$ and $\frac{d\tilde{\lambda}(\tilde{\mu}_f(\tilde{r}_f))}{d\mu} < 0$. As $\frac{\tilde{r}_f}{\tilde{r}_b} \leq 1-\delta(\tilde{\mu}_f(\tilde{r}_f))$,

$$\left.\frac{d\Pi_b(\mu)}{d\mu}\right|_{\mu=\tilde{\mu}_f(\tilde{r}_f)} = \left(\tilde{r}_b - \frac{\tilde{r}_f}{1-\delta(\tilde{\mu}_f(\tilde{r}_f))}\right)\left.\frac{d\tilde{\lambda}(\mu)}{d\mu}\right|_{\mu=\tilde{\mu}_f(\tilde{r}_f)} - \frac{\tilde{r}_f\delta'(\tilde{\mu}_f(\tilde{r}_f))\tilde{\lambda}(\tilde{r}_f)}{(1-\delta(\tilde{\mu}_f(\tilde{r}_f)))^2} < 0.$$

As $\Pi_b(\mu)$ is unimodal in μ (Proposition 1), this implies that $\tilde{\mu}_f(\tilde{r}_f) > \tilde{\mu}_b(\tilde{r}_b)$, which leads to a contradiction. Thus, $\tilde{\lambda}(\tilde{r}_f) < \tilde{\lambda}(\tilde{r}_b)$ and $\tilde{\mu}_f(\tilde{r}_f) > \tilde{\mu}_b(\tilde{r}_b)$. As $\delta(\mu)$ is increasing in μ, $\delta(\tilde{r}_f) > \delta(\tilde{r}_b)$. From (15), we have

$$S(\tilde{r}_f) = -\beta(\Lambda - \tilde{\lambda}(\tilde{r}_f)) < -\beta(\Lambda - \tilde{\lambda}(\tilde{r}_b)) = S(\tilde{r}_b).$$

As

$$W(\mu) = \frac{R(1-\delta(\mu))}{\theta} - \frac{t}{\theta};\ T(\mu) \ = \ \frac{R}{\theta} - \frac{t}{\theta(1-\delta(\mu))}, \tag{32}$$

it can be easily shown that both $W(\mu)$ and $T(\mu)$ are decreasing in μ. Because $\tilde{\mu}_f(\tilde{r}_f) > \tilde{\mu}_b(\tilde{r}_b) > \mu^*$, $W(\tilde{r}_f) < W(\tilde{r}_b) < W(\mu^*)$ and $T(\tilde{r}_f) < T(\tilde{r}_b) < T(\mu^*)$. □

Proof of Proposition 5. Differentiating $\Pi_f(\mu)$ given in (19) with respect to μ yields

$$\frac{d\Pi_f(\mu)}{d\mu} = \frac{c}{\mu^2}\frac{\Lambda}{1-\delta(\mu)} + \left(r_f - \frac{c}{\mu}\right)\frac{\Lambda\delta'(\mu)}{(1-\delta(\mu))^2} > 0.$$

Thus, $\Pi_f(\mu)$ is increasing in μ. The HCP will select the largest service rate that satisfies the full coverage requirement (i.e., $\max\{\mu|\tilde{\lambda}(\mu) \geq \Lambda\}$).

Differentiating $\Pi_b(\mu)$ given in (20) with respect to μ yields

$$\frac{d\Pi_b(\mu)}{d\mu} = \frac{co^{'}(\mu)}{o^2(\mu)}\Lambda,$$

which equals zero at $\mu = \mu^o$. According to Lemma 1, $o^{''}(\mu^o) < 0$. Thus,

$$\left.\frac{d^2\Pi_b(\mu)}{d\mu^2}\right|_{\mu=\mu^o} = \frac{co^{''}(\mu^o)}{o^2(\mu^o)}\Lambda < 0.$$

That is, $\Pi_b(\mu)$ is unimodal in μ. Therefore, when μ^o satisfies the full coverage requirement (i.e., $\lambda(\mu^o) \geq \Lambda$), the HCP's optimal service rate under the full coverage equals μ^o. Otherwise, if $\tilde{\lambda}(\mu^o) < \Lambda$, the full coverage requirement $\tilde{\lambda}(\mu) \geq \Lambda$ should be binding. Since $\tilde{\lambda}(\mu)$ is unimodal in μ, there exist two solutions that satisfy $\tilde{\lambda}(\mu) = \Lambda$. We denote these two solutions as $\underline{\mu}$ and $\bar{\mu}$, respectively. Without loss of generality, we assume that $\underline{\mu} < \bar{\mu}$. Taking the derivative of the second term of $\tilde{\lambda}(\mu)$ in (8) with respect to μ, we have

$$\frac{d}{d\mu}\left(\frac{\theta(1-\delta(\mu))}{R(1-\delta(\mu))-t}\right) = \frac{t\theta\delta^{'}(\mu)}{(R(1-\delta(\mu))-t)^2} > 0.$$

Therefore,

$$\frac{\theta(1-\delta(\underline{\mu}))}{R(1-\delta(\underline{\mu}))-t} < \frac{\theta(1-\delta(\bar{\mu}))}{R(1-\delta(\bar{\mu}))-t}.$$

Because $\tilde{\lambda}(\underline{\mu}) = \tilde{\lambda}(\bar{\mu}) = \Lambda$, $o(\underline{\mu}) < o(\bar{\mu})$. Since $\Pi_b(\mu)$ is increasing in $o(\mu)$, $\Pi_b(\underline{\mu}) < \Pi_b(\bar{\mu})$. Thus, when $\tilde{\lambda}(\mu^o) \leq \Lambda$, the HCP's optimal service rate under the full coverage equals $\bar{\mu} = \max\{\mu|\tilde{\lambda}(\mu) \geq \Lambda\}$. □

Proof of Proposition 6. Differentiating $S(\mu)$ given in (21) with respect to μ yields

$$\frac{dS(\mu)}{d\mu} = -\frac{t\delta^{'}(\mu)}{(1-\delta(\mu))^2} + \frac{\theta o^{'}(\mu)}{(o(\mu)-\Lambda)^2}. \tag{33}$$

Because $S(\mu)$ achieves its maximum at μ_u, μ_u should satisfy the FOC of $S(\mu)$, which further implies that $o^{'}(\mu_u) > 0$. From Lemma 1, $\mu_u < \mu^o$ and $o^{''}(\mu_u) < 0$. By using (23), $\delta^{''}(\mu)(1-\delta(\mu)) + (\delta^{'}(\mu))^2 > 0$. Thus,

$$\left.\frac{d^2S(\mu)}{d\mu^2}\right|_{\mu=\mu_u} = -\frac{t[\delta^{''}(\mu_u)(1-\delta(\mu_u)) + 2(\delta^{'}(\mu_u))^2]}{(1-\delta(\mu_u))^3} + \frac{\theta o^{''}(\mu_u)}{(o(\mu_u)-\Lambda)^2} - \frac{\theta(o^{'}(\mu_u))^2}{(o(\mu_u)-\Lambda)^2} < 0,$$

which shows that $S(\mu)$ is unimodal in μ. Therefore, when $B \geq c\cdot\Lambda/o(\mu_u)$, μ_u is feasible such that $\mu^* = \mu_u$. While when $B < c\cdot\Lambda/o(\mu_u)$, the budget constraint (22) is binding. As $o(\mu)$ is unimodal in μ, there exists two service rates that make the budget constraint (22) binding; that is, $o(\mu) = c\cdot\Lambda/B$. From (7), because $\delta(\mu)$ is increasing in μ, we can easily know that

the optimal service rate is the smaller one. Finally, to facilitate our analysis in Corollary 4, we next show that when $B < c \cdot \Lambda / o(\mu_u)$, $\mu_u < \mu^*$. If $B < c \cdot \Lambda / o(\mu_u)$, then $o(\mu_u) < o(\mu^*)$. Since $T(\Lambda, \mu)$ is decreasing in $o(\mu)$, $T(\Lambda, \mu_u) > T(\Lambda, \mu^*)$. In this case, if $\mu_u \geq \mu^*$ such that $\delta(\mu_u) \geq \delta(\mu^*)$, then $U(\Lambda, \mu_u) < U(\Lambda, \mu^*)$, which contradicts the definition of μ_u. Hence, when $B < c \cdot \Lambda / o(\mu_u)$, $\mu_u < \mu^*$. □

Proof of Corollary 4. For ease of exposition, we denote $\mu_l = \max\{\mu | \tilde{\lambda}(\mu) \geq \Lambda\}$. According to Proposition 5, when $\tilde{\lambda}(\mu^o) \leq \tilde{\lambda}(\tilde{r}_f)$ and $\tilde{\lambda}(\mu^o) \leq \Lambda \leq \tilde{\lambda}(\tilde{r}_f)$, $\tilde{\mu}_f = \tilde{\mu}_b = \mu_l$. While when $\Lambda < \min\{\tilde{\lambda}(\mu^o), \tilde{\lambda}(\tilde{r}_f)\}$, $\tilde{\mu}_b = \mu^o$ and $\tilde{\mu}_f = \mu_l$. Since μ_l is the larger solution of $\tilde{\lambda}(\mu) = \Lambda$ and $\Lambda < \tilde{\lambda}(\mu^o)$, $\tilde{\mu}_b = \mu^o < \tilde{\mu}_f$. And from (6) we have that the total waiting time T is decreasing in $o(\mu)$. As $o(\mu)$ is maximized at μ^o, $T(\Lambda, \tilde{\mu}_b) < T(\Lambda, \tilde{\mu}_f)$. In addition, because $\delta(\mu)$ is increasing in μ, $\delta(\tilde{\mu}_f) > \delta(\tilde{\mu}_b)$.

Next, we can show that $n(\tilde{\mu}_b) < n(\tilde{\mu}_f)$ as $n(\mu)$ is increasing in μ. As $n(\tilde{\mu}_b) < n(\tilde{\mu}_f)$ and $T(\Lambda, \tilde{\mu}_b) < T(\Lambda, \tilde{\mu}_f)$, we can show the following relationship regarding the patient utility:

$$U(\Lambda, \tilde{\mu}_f) = R - [n(\tilde{\mu}_f) \cdot t + \theta \cdot T(\Lambda, \tilde{\mu}_f)] < R - [n(\tilde{\mu}_b) \cdot t + \theta \cdot T(\Lambda, \tilde{\mu}_b)] = U(\Lambda, \tilde{\mu}_b).$$

Consequently, $S(\tilde{\mu}_f) = \Lambda \cdot U(\Lambda, \tilde{\mu}_f) < \Lambda \cdot U(\Lambda, \tilde{\mu}_b) = S(\tilde{\mu}_b)$.

We then compare the centralized healthcare system with the FFS and BP schemes. Since the optimal service rate under the centralized healthcare system should satisfy the full coverage requirement (i.e., $\tilde{\lambda}(\mu^*) > \Lambda$), $\mu^* < \mu_l$. Furthermore, we have shown in Proposition 6 that when $B \geq c \cdot \Lambda / o(\mu_u)$, $\mu^* < \mu^o$; while when $B < c \cdot \Lambda / o(\mu_u)$, μ^* is the smaller root that solves $o(\mu^*) = c \cdot \Lambda / B$. As $o(\mu)$ is unimodal in μ and achieves its maximum at μ^o, the inequality $\mu^* < \mu^o$ still holds when $B < c \cdot \Lambda / o(\mu_u)$. Therefore, $\mu^* < \tilde{\mu}_b \leq \tilde{\mu}_f$. Moreover, because $\delta(\mu)$ is increasing in μ, $\delta(\mu^*) < \delta(\tilde{\mu}_b) \leq \delta(\tilde{\mu}_f)$. We have shown in Proposition 6 that $\mu_u \leq \mu^*$; therefore, $\mu_u \leq \mu^* < \tilde{\mu}_b \leq \tilde{\mu}_f$. Combining this with the fact that $S(\mu)$ given in 21 is unimodal in μ, we get $S(\mu^*) > S(\tilde{\mu}_b) \geq S(\tilde{\mu}_f)$. When $\tilde{\lambda}(\mu^o) \leq \Lambda \leq \tilde{\lambda}(\tilde{r}_f)$, $\tilde{\mu}_f = \tilde{\mu}_b = \mu_l$. Because $\tilde{\lambda}(\mu)$ is unimodal in μ, and both μ^o and μ_l are at the right-hand side of the mode of $\tilde{\lambda}(\mu)$, $\mu_l < \mu^o$. By using the unimodality of $o(\mu)$, we have $o(\mu^*) < o(\mu_l)$. Because $T(\Lambda, \mu)$ is decreasing in $o(\mu)$, $T(\Lambda, \mu^*) > T(\Lambda, \mu_l) = T(\Lambda, \tilde{\mu}_f) = T(\Lambda, \tilde{\mu}_b)$. When $\Lambda < \min\{\tilde{\lambda}(\mu^o), \tilde{\lambda}(\tilde{r}_f)\}$, $\tilde{\mu}_b = \mu^o$ and $\tilde{\mu}_f = \mu_l$. As $o(\mu)$ achieves its maximum at μ^o and $T(\Lambda, \mu)$ is decreasing in $o(\mu)$, the inequality $T(\Lambda, \mu^*) > T(\Lambda, \tilde{\mu}_b)$ still holds. □