The Impact of Reimbursement Policy on Patient Welfare, Readmission 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

Faculty of Business, the Hong Kong Polytechnic University, Hong Kong; and
School of Economics and Management, Southwest Jiaotong University, Chengdu, China, lighting.zhao@connect.polyu.hk

Abstract

This paper examines the impact of two reimbursement schemes on patient welfare, readmission rate, and waiting time in a three tiered public healthcare system comprising (a) a public funder who decides on the reimbursement rate to maximize patient welfare, (b) a public healthcare provider (HCP) who decides on the service rate (which affects readmission rate and operating cost), and (c) a pool of (waiting time sensitive) patients who decide whether or not to seek elective treatments. We focus our analysis on (1) a Fee-For-Service (FFS) scheme under which the HCP receives payment each time a patient is admitted (or readmitted); and (2) a Bundled Payment (BP) scheme under which the HCP receives a lump sum payment for the entire episode of care for each patient (regardless of the number of readmissions). By considering an M/M/1 queueing model with endogenous arrivals and readmissions, we analyze a three-stage Stackelberg game to determine the patient's initial admission rate, the HCP's service rate (which affects the readmission rate), and the funder's reimbursement rate. This analysis enables us to compare the equilibrium outcomes (patient welfare, readmission rate and waiting time) associated with the FFS and BP schemes. We find that, when the patient pool is large, the BP scheme dominates in terms of higher patient welfare and lower readmission rate, but the FFS scheme dominates in terms of waiting time. However, when the patient pool is small, the BP scheme dominates the FFS scheme in all three performance measures.

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

1 Introduction

Public healthcare systems are facing many challenges: operating cost escalates, service quality deteriorates, and waiting time lengthens. 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 of 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 UK, waiting times for elective surgery are considered by the public as 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 surgeries, and 80.7 days for hernia repairs (Hurst and Sicilliani 2003).[1] In Hong Kong, the waiting time for cataract surgery is longer than eight months.[2] Because excessive long waiting time causes patient dissatisfaction, some public healthcare systems (such as the UK) include waiting time (especially for elective surgeries) as a key performance measure (Dimakou 2013), and others (such as South Australia) are committed to reduce waiting time.[3]

Many healthcare professionals believe that an effective reimbursement scheme can entice healthcare providers (HCPs) to reduce waiting time, contain cost and improve service quality. Currently, the predominant scheme is called Fee-For-Service (FFS) under which a 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, higher readmissions will ensue (Kociol et al. 2012); and the FFS scheme creates major concerns including: (1) excessive treatments (Davis 2007); (2) high readmissions (Fenter and Lewis 2008); and (3) low service quality at high cost (Calsyn and Lee 2012).

To improve service quality and contain cost, the Centers for Medicare and Medicaid Services (CMS) in the United States is shifting gradually 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 survey study claimed that, relative to FFS, the BP reimbursement scheme can reduce the cost per episode of care by 3% (Japsen 2015).

[1] See "NHS patients waiting longer for routine operations under coalition" at http://www.theguardian.com/society/2014/jul/04/nhs-patients-waiting-longer-for-routine-operations-under-coalition.

[2] See "Waiting Time for Cataract Surgery", released on the Hong Kong government website at http://www.ha.org.hk/visitor/ha_visitor_text_index.asp?Parent_ID=214172&Content_ID=214184.

[3] See "Elective surgery services", posted on the South Australian Government website at http://www.sahealth.sa.gov.au/wps/wcm/connect/Public+Content/SA+Health+Internet/Health+services /Elective+surgery+services/.

At the same time, Ontario (Canada) is examining the effectiveness of the BP scheme since 2011[4]; and the Australian government is considering the BP scheme in 2015[5].

While the BP scheme has been adopted by some public healthcare systems, many public systems continue to operate under the FFS scheme because the underlying implications are not well understood. Therefore, it is important to gain a deeper understanding about the implications of these two schemes on certain performance measures including patient welfare and service quality (readmissions and waiting time). In this paper, we compare these performance measures associated with the FFS and BP schemes for providing outpatient elective care services in a public healthcare system that consists of a funder, a 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. Specifically, in our model, the funder acts as the first leader who determines the reimbursement rate to maximize the patient welfare. Given the reimbursement rate, the HCP acts as the second leader who decides on the service rate to maximize its profit, where a higher service rate yields a higher readmission 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' admissions into consideration. Hence, the patient's admission rate is *endogenously* determined according to a Nash equilibrium.[6]

Embedded in our three-stage Stackelberg game is an M/M/1 queueing model with endogenous patient arrivals and readmissions. This queueing model enables us to determine the service rate (decided by the HCP) and the corresponding patient arrival rate, readmission rate, waiting time and patient welfare for any given reimbursement rate (specified by the funder) under both FFS and BP schemes. By comparing the equilibrium outcomes (the patient welfare, readmission 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 patient welfare and lower readmission rate. However, the

[4] See "Ontario Funds Bundled Care Teams to Improve Patient Experience" at https://news.ontario.ca/mohltc/en/2015/09/ontario-funds-bundled-care-teams-to-improve-patient-experience.html.

[5] See the report of the 2015 Primary Health Care Advisory Group, released on the Australian government website at http://www.health.gov.au/internet/main/publishing.nsf/Content/primary-phcag-report.

[6] 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 (EU) has 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).

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 welfare, lower readmission 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 admissions and random readmissions 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 patient welfare, readmission rate and waiting time.

This paper is organized as follows. §2 reviews the relevant literature. In §3, we present our queuing model and establish some preliminary results. In §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 §5, we compare the equilibrium outcomes under two schemes when the patient population is small. Concluding remarks are provided in §6. All proofs are relegated to the online Appendix A.

2 Literature Review

This paper is 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 patient's health. By using a dynamic principal-agent game theoretic 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. Other research papers in this stream include Jiang et al. (2012), Ata et al. (2013) and Bavafa et al. (2013).

As public funders in different countries are contemplating whether they shall change the payment scheme from FFS to BP, researchers are developing different models to compare the performance measures associated with these two schemes. The first paper in this area is by Adida et al. (2014). They 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 analysing a two-stage model, they examine the impact of the FFS and BP schemes on patient selection and treatment intensity. They find that, due to risk aversion of the HCP, 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, they propose two alternative payment systems that may induce system optimal decisions. Next, 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 readmission 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 readmission-reduction efforts. Finally, 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 proposal 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. (2014) in which the HCP selects which type of patients to admit, we consider a situation in which patients are sensitive to waiting time and they can elect not to seek elective care from the public HCP so that the patient's arrival rate is endogenously determined by the patients (not the HCP). Second, unlike those two-stage models developed by Adida et al. (2014) and Andritsos and Tang (2015) in which a patient can only be readmitted at most once, we use a queueing model with random readmissions 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 patient welfare, readmission rate, and waiting times associated with FFS and BP.

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 when the arrival rate is endogenously determined by the customers, Anand et al. (2011) investigate the optimal pricing strategy for the service provider. While Anand et al. (2011) find that a lower service rate will increase both the waiting time and the service quality, it is interesting to note that this finding continues to hold in our model when the patient population is large. However, due to the fact that there are two inter-related customer arrival streams (initial admissions and readmissions) in our model, we obtain a different result when the patient population is so small that all patients will seek (initial) admissions. In this case, a (slightly) lower service rate will not affect the initial admission rate; however, it will reduce the waiting time due to a lower readmission rate (i.e., higher service quality).

Along 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 speed-quality 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 health care 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 readmission rate is endogenously determined by the HCP (via its selection of service rate). Also, 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

Consider a public healthcare system consisting of a funder who sets the reimbursement rate subject to a limited budget, a 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[7]. 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 readmissions (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. However, due to balking patients, the initial arrival rate to the HCP is exogenously given by λ in the base model. (However, we shall examine the case when the initial arrival rate to the HCP is endogenously determined by the patients

[7] 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).

in §4.1.)

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.[8] Akin to Anand et al. (2011), we shall assume that the HCP's manpower capacity (i.e., the number of doctors) is fixed.[9] However, the HCP can change its service rate μ by adjusting the service time per patient. To capture the speed-quality trade-off, we shall consider the case in which the patient readmission is more likely to occur when the HCP increases its service rate.[10]

After discharge, the patient is either cured or readmitted with probability $\delta(\mu)$. For tractability, we make the following assumptions about the readmission rate $\delta(\mu)$:

Assumption 1: The readmission 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; i.e., $\log(1 - \delta(\mu))$ is concave so that $g(\mu) = \delta^{'}(\mu)/(1 - \delta(\mu))$ is increasing in μ.

Assumption 1 captures an empirical fact that the readmission rate is increasing in the service rate μ (Kociol et al. 2012). 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 $b > 0$ satisfy both assumptions, where the logistic function is a standard approach to measure the relationship between the readmission rate and other variables in the healthcare management literature (e.g., Fethke et al. 1986, Morrow-Howell and Proctor 1993).

Based on above assumptions, we can model the healthcare system as an M/M/1 queue with random readmissions (via Bernoulli trials) as depicted in Figure 1. Note that the service rate for new patients and that for readmitted patients are assumed to be the same. This assumption is reasonable for outpatient elective care service (such as hernia repair operations) where the appointment block for each patient is normally fixed, regardless of whether the patient is new or readmitted.[11]

[8] 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 to a hospital intensive care unit follows a Poisson process, and the service time follows an exponential distribution.

[9] 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, remains stable from 2002 to 2012 (Hing and Hsiao 2014).

[10] Kociol et al. (2012) find empirical evidence that the readmission rate is increasing in the service rate μ.

[11] For example, the appointments with primary care doctors are normally scheduled at 15-minute intervals in the United States (see https://www.washingtonpost.com/opinions/when-medical-care-is-delivered-in-15-minute-doses-theres-not-much-time-for-caring/2015/11/13/85ddba3a-818f-11e5-a7ca-

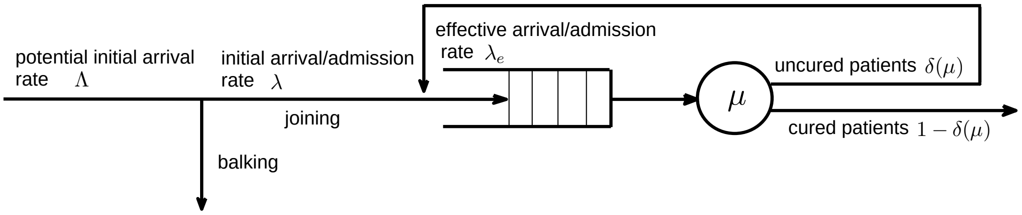

effective arrival/admission
rate λ_e
potential initial arrival
rate Λ
initial arrival/admission
rate λ
joining
balking
μ
uncured patients $\delta(\mu)$
cured patients $1-\delta(\mu)$

Figure 1: A Schematic of the Model

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

Lemma 1. *The cure service rate $o(\mu)$ is quasi-concave in μ. The optimal μ^o is attained when the elasticity of the cure rate equals* 1*; i.e., when*

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

Also, 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 $\mu\cdot g(\mu)<1$, i.e., 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 $\mu\cdot g(\mu)=1$.

6ab6ec20f839_story.html). In the United Kingdom, the physicians allocate almost the same amount of time for the initial visits (i.e., slightly less than 11 minutes) and the follow-up ones (i.e., slightly less than 10 minutes) (Konrad et al. 2010).

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 admission and the waiting time of all potential subsequent readmissions during a *medical episode*. Let λ and λ_e denote the patients' initial arrival rate (i.e., the arrival rate of newly admitted patients)[12] and the patients' effective arrival rate (that includes initial admissions and all subsequent readmissions), respectively. 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 readmissions (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 $L = \frac{\lambda_e}{\mu - \lambda_e} = \frac{\lambda}{o(\mu) - \lambda}$. Let W and T denote the expected waiting time per visit and the expected total waiting time per medical episode; respectively. By using the Little's law, we have $L = \lambda_e \cdot W = \lambda \cdot T$. By combining these two observations, we get:

$$\begin{aligned} W(\lambda, \mu) &= \frac{1}{\mu - \lambda_e} = \frac{1 - \delta(\mu)}{o(\mu) - \lambda}, && (5) \\ T(\lambda, \mu) &= \frac{1}{o(\mu) - \lambda}. && (6) \end{aligned}$$

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 $o(\mu)$.

[12] 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 admission rate" throughout this paper.

3.3 Patient Utility

For any given service rate μ, a waiting time sensitive patient who seeks admission from the HCP derives her utility $U(\lambda,\, \mu)$, where

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

in which $n(\mu)$ is the number of admissions 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 *non-pecuniary disutility* associated with each admission, and θ is the imputed cost associated with waiting. Here, we assume that the waiting time will not cause adverse effects (i.e., worsening patients' symptoms). This assumption is reasonable for elective surgeries and it is supported by the empirical evidence established by Hurst and Siciliani (2003).

Knowing the readmission rate and waiting time,[13] each patient will seek admission if and only if the utility associated with the admission $U(\lambda,\, \mu) \geq 0$.[14] By using the fact that the utility $U(\lambda,\, \mu)$ given in (7) is strictly decreasing in λ, we can determine the initial admission rate $\tilde{\lambda}(\mu)$ when admissions are endogenously decided by the patients. First, consider the case when the potential initial admission 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 admission 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 admission rate Λ is sufficiently small so that $U(\Lambda, \mu) \geq 0$. In this case, all potential patients will seek admissions so that $\tilde{\lambda} = \Lambda$. We shall refer to this case as the *full coverage* scenario. To avoid repetition and to ease our 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, and provide similar analysis for the case where patients are fully covered under one scheme but are partially covered under the other scheme in the online Appendix B.

So far, we have established the relationships among the initial admission rate λ, the readmission rate $\delta(\mu)$, the patient utility $U(.,.)$, the patient's waiting time per visit $W(.,.)$,

[13] In many countries, the readmission rate and waiting time are common knowledge. For example, the Australian government releases the Australian hospital statistics report at http://www.aihw.gov.au/publication-detail/?id=60129553174, in which the information regarding the waiting time and the readmission rate for the elective surgery can be found on page 35 and page 52, respectively.

[14] When the imputed disutility associated with each admission 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 admission. 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 admission in equilibrium.

and the patient's 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 (the government or a private insurer) first selects the payment scheme (FFS or BP) and the reimbursement rate. Anticipating the funder's reimbursement rate, 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 overcrowding public healthcare systems with long waiting time. 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 admission based on her expected utility. Anticipating the patients' initial admission rate and effective admission rate in equilibrium $\tilde{\lambda}(\mu)$ and $\tilde{\lambda}_e(\mu)$, we shall derive the HCP's service rate decisions and the funder's reimbursement decisions under 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 admission 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.[15]

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. 2. The HCP determines the optimal service rate μ_f. 3. The patients decide to join or balk and the initial admission rate in equilibrium is equal to $\tilde{\lambda}(\mu_f)$.	1. The funder determines the optimal reimbursement rate r_b for each episode. 2. The HCP determines the optimal service rate μ_b. 3. The patients decide to join or balk and the initial admission rate in equilibrium is equal to $\tilde{\lambda}(\mu_b)$.

[15] Under the FFS scheme, the HCP receives 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.

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 admission 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 admission, 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 admission 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 (5) 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 $R(1-\delta(\mu)) > t$.

In the next section, we shall utilize the initial admission rate $\tilde{\lambda}(\mu)$ and the effective admission 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., $U(\Lambda, \mu) < 0$), in equilibrium both the initial admission rate $\tilde{\lambda}(\mu)$ and the effective admission rate $\tilde{\lambda}_e(\mu)$ are unimodal in μ. Moreover, the mode of $\tilde{\lambda}(\mu)$ is smaller than that of $\tilde{\lambda}_e(\mu)$.*

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 admissions (Rabin 2014). On the other hand, a higher service rate will cause a higher readmission rate, which discourages patients from seeking admissions.[16] Since 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. Hence, $\tilde{\lambda}(\mu)$ and $\tilde{\lambda}_e(\mu)$ are unimodal in μ as depicted in Figure 2.

[16] The readmission rate can affect a patient's joining-or-balking decision. For example, Varkevisser et al. (2012) show that patients prefer hospitals with low readmission rates and a 1% reduction in the readmission rate is associated with a 12% increase in hospital demand.

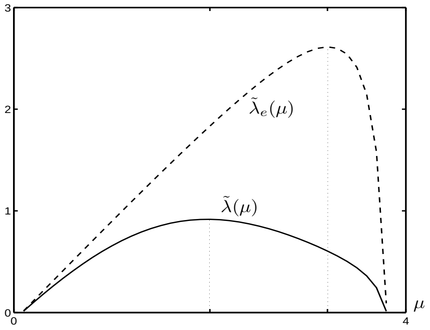


Figure 2: Initial and effective admission rates $\big(\delta(\mu) = \frac{1}{1+e^{-\mu+2}}$, $\theta = 0.5$, $t = 1$, $R = 8\big)$

Corollary 1 (along with Figure 2) has two implications. First, when μ is large, having the physicians to work faster can discourage patients to seek admissions (i.e., both $\tilde{\lambda}(\mu)$ and $\tilde{\lambda}_e(\mu)$ will decrease) since the readmission 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 admission rate $\tilde{\lambda}(\mu)$ but it can increase the effective admission rate $\tilde{\lambda}_e(\mu)$ (due to the significant increase in the readmission rate $\delta(\mu)$). Therefore, when choosing the service rate μ under different payment schemes, the HCP should take into account the impact of μ on the initial admission rate and the effective admission rate in equilibrium. 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 admission rate $\tilde{\lambda}(\mu)$ and the equilibrium effective admission 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 admission under the FFS scheme and r_b for each episode under the BP scheme. Hence, 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 is mainly attributed to the length of patients' outpatient visits (i.e., personnel, nurse, consumable items, etc.) 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. Combining these observations, we can formulate the HCP's problem under the two schemes as

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

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

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 $o(\mu)$, i.e., $\tilde{\mu}_b(r_b) < \mu^o$. 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 substituting the optimal $\tilde{\mu}_s(r_s)$ (given in (12) and (13) respectively) into $\delta(\mu)$, $\tilde{\lambda}(\mu)$ given in (8), $\tilde{\lambda}_e(\mu)$ given in (9), and $\Pi_s(\mu)$ given in (10) and (11), we can express these quantities as functions of the reimbursement rate r_s so that $\delta(r_s) = \delta(\tilde{\mu}_s(r_s))$, $\tilde{\lambda}(r_s) = \tilde{\lambda}(\tilde{\mu}_s(r_s))$, and $\Pi_s(r_s) = \Pi_s(\tilde{\mu}_s(r_s))$. (For ease of exposition, we shall suppress the arguments when convenient.) We shall use these quantities to determine the funder's reimbursement decision r_s, $s = f, b$, under the two schemes. In preparation, let us examine the properties of these quantities 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 readmission rate $\delta(r_s)$ are decreasing in r_s.*

2. *The initial admission 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 on treating each patient and therefore, less patients are readmitted to 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 admissions (i.e., to increase the effective admission rate $\tilde{\lambda}_e(\mu)$ under the FFS scheme and to increase the initial admission rate $\tilde{\lambda}(\mu)$ under the BP scheme). Having this desire in mind, it is easy to observe from (12) and (13) that, in order to increase $\tilde{\lambda}_e(\mu)$ under FFS and $\tilde{\lambda}(\mu)$ under BP, the HCP has to lower its service rate μ. This explains the first statement.

Next, the second statement informs us that a higher reimbursement rate will encourage more patients to seek admission (i.e., the initial admission 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 admission rate $\tilde{\lambda}_e(\mu)$ under FFS and increase the initial admission rate $\tilde{\lambda}(\mu)$ under BP. Also, with a lower service rate and a higher effective admission rate $\tilde{\lambda}_e(\mu)$ under FFS, it is easy to check from (3) that the initial admission 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 admission rate and the effective admission 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 higher admission rate and higher reimbursement rate under both schemes.

In summary, Corollary 2 highlights the trade-off between readmission rate and 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 patient welfare $S(r_s)$. When the patient initial admission rate is endogenous, the patient welfare generally consists of two components:

(a) the utility of those patients who seek admissions; and (b) the penalty cost associated with those patients who balk (without seeking admissions). As stated in the introduction, it is common that patients who demand for public care abandon and seek help elsewhere due to the long waiting time. Therefore, accessibility is central to the performance of the public healthcare systems (Levesque et al. 2013, and Adritsos and Tang, 2014). To incorporate accessibility into the funder's objective, we impose a penalty cost β for each balking patient.

Under the partial coverage case, we know that $\tilde{\lambda}(r_s)$ is the initial admission 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 a disutility β, the patient welfare $S(r_s) = \tilde{\lambda}(r_s) \cdot U(\tilde{\lambda}(r_s), \tilde{\mu}_s(r_s)) - \beta(\Lambda - \tilde{\lambda}(r_s))$. By taking the funder's budget B and the HCP's participation constraint into consideration, we can formulate the funder's problem as follows:

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

$$s.t. \qquad \textit{Budget constraint}: \begin{cases} (FFS)\ r_f \cdot \tilde{\lambda}_e(r_f) \leq B, \\ (BP)\ r_b \cdot \tilde{\lambda}(r_b) \leq B, \end{cases} \tag{15}$$

$$\textit{Participation constraint}: \Pi_s(r_s) \geq 0.$$

Recall from §4.1 that $U(\tilde{\lambda}(r_s), \tilde{\mu}_s(r_s)) = 0$ under the partial coverage case, it is easy to check from (14) that the funder's objective is equivalent to maximizing the initial admission rate $\tilde{\lambda}(r_s)$.[17] 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 patient welfare $S(r_s)$ given in* (14) *is increasing in r_s. The funder's budget constraint (15) 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 $\tilde{r}_b \cdot \tilde{\lambda}_e(\tilde{r}_b) = B$ 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 FFS and BP schemes (Donnelly and Swinford 2013).[18] We can further establish the following comparative statics 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 readmission rate $\delta(\tilde{r}_s)$ are decreasing in B. However, the optimal*

[17] Under the full coverage case, each patient can obtain a positive utility (i.e., $U(\Lambda, \mu) > 0$). Therefore, the funder's objective is no longer equivalent to maximizing the initial admission rate.

[18] See "NHS is about to run out of cash top official warns" at http://www.telegraph.co.uk/news/health/news/10162848/NHS-is-about-to-run-out-of-cash-top-official-warns.html.

reimbursement rate $\tilde{r}_s$, the initial admission 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 as 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 the patient access and reduce the readmission 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 v.s. BP

So far, we have derived the equilibrium outcomes of different performance metrics (the patient welfare, initial admission rate, service rate, readmission rate and waiting times) associated with both payment schemes under the partial coverage scenario as presented in Propositions 1 and 2. We now compare these performance outcomes between the FFS scheme and the BP scheme. Specifically, our comparison yields the following results.

Proposition 3. *Consider the case in which potential patients are partially covered under both FFS and BP schemes. Then, relative to the FFS scheme,*

1. *both the patient welfare and the initial admission rate are higher under the BP scheme (i.e., $S(\tilde{r}_b) > S(\tilde{r}_f)$ and $\tilde{\lambda}(\tilde{r}_b) > \tilde{\lambda}(\tilde{r}_f)$);*

2. *both the service rate and the readmission rate are lower under the BP scheme (i.e., $\tilde{\mu}_b(\tilde{r}_b) < \tilde{\mu}_f(\tilde{r}_f)$ and $\delta(\tilde{r}_b) < \delta(\tilde{r}_f)$); and*

3. *both the waiting time per visit and the total waiting time are higher under the BP scheme (i.e., $W(\tilde{r}_b) > W(\tilde{r}_f)$ and $T(\tilde{r}_b) > T(\tilde{r}_f)$).*

Proposition 3 implies that when potential patients are partially covered under both FFS and BP schemes, the BP scheme dominates the FFS scheme in terms of the patient welfare and service quality, but the FFS scheme outperforms the BP scheme in terms of the waiting time. 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 is readmitted to the system. Hence, the HCP under the BP scheme has incentives to reduce the service rate so as to reduce the readmission rate. However, reducing the readmission rate attracts higher initial admission rate under the BP scheme. Furthermore, from (7), if the potential patients are partially covered, then the initial admission rate solves $U(\tilde{\lambda}, \mu) = 0$. Due to a lower readmission 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 patient welfare is equivalent to maximizing the initial admission rate under the partial coverage scenario, the patient welfare under the BP scheme is also larger.

Proposition 3 is consistent with the findings of previous studies showing that the FFS scheme is effective for reducing the waiting time but not for improving service quality. For example, Blomqvist and Busby (2013) show that the FFS scheme is effective for 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 the elective surgery.

In summary, by 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 for improving the patient welfare and reducing the readmission rate; however, the FFS scheme is more effective for 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 3 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 FFS and BP schemes. This case is suitable for the public HCP in the rural area, which normally has a low patient volume (DiChiara 2015).[19] Recall that under the full coverage scenario, $U(\Lambda, \mu) \geq 0$ (see §4.1). This condition can be further simplified as $\tilde{\lambda}(\mu) \geq \Lambda$ after some algebra, where $\tilde{\lambda}(\mu)$ is given in (8). In other words, when the condition under the partial coverage scenario is infeasible (i.e., $\tilde{\lambda}(\mu) \geq \Lambda$), the healthcare system will achieve the full coverage scenario.

Under the full coverage scenario, the initial admission rate is given by Λ and the effective admission 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 problems under the full coverage

[19] See "Rural Hospitals Address Medicare Reimbursement Cut Concerns" at http://revcycleintelligence.com/news/rural-hospitals-address-medicare-reimbursement-cut-concerns.

scenario as follows:

$$\begin{aligned}(FFS)\ \max_{\mu}\ \Pi_f(\mu) &= \left(r_f - \frac{c}{\mu}\right)\frac{\Lambda}{1-\delta(\mu)}, && (16)\\ s.t. \quad & \tilde{\lambda}(\mu) \geq \Lambda, \\ (BP)\ \max_{\mu}\ \Pi_b(\mu) &= \left(r_b - \frac{c}{o(\mu)}\right)\Lambda, && (17)\\ s.t. \quad & \tilde{\lambda}(\mu) \geq \Lambda,\end{aligned}$$

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

Proposition 4. *Consider the case in which all potential patients are fully covered under both 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 satisfies $\tilde{\mu}_f = \max\{\mu,$ 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 4 is as follows. First, observe from (16) that, the HCP's variable cost c/μ decreases in μ while its effective arrival rate $\Lambda/(1-\delta(\mu))$ increases in μ. Hence, the HCP's profit $\Pi_f(\mu)$ under the FFS scheme is always increasing in μ. These observations imply that the HCP will choose the largest service rate in equilibrium that ensures that 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 (17) 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 (17) 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. Whereas, 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 that the potential patients are fully covered (i.e., $\max\{\mu : \tilde{\lambda}(\mu) \geq \Lambda\}$).

Finally, unlike the partial coverage scenario, Proposition 4 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 of the performance metrics that we are going to compare such as the patient welfare, the initial admission rate and the total waiting time only depend on the service rate and the initial admission rate, for ease of exposition, we thus omit the analysis of the funder's reimbursement rate decisions under the full coverage case. (See the online Appendix B for details.)

To guarantee that the healthcare system in equilibrium achieves the full coverage, 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 admission rate under the partial coverage case. According to Proposition 3, this condition is equivalent to $\Lambda \leq \tilde{\lambda}(\tilde{r}_f)$. By comparing different performance metrics associated with the FFS and BP schemes under the full coverage (i.e., $\Lambda \leq \tilde{\lambda}(\tilde{r}_f)$), we get the following results.

Corollary 4. *Suppose that potential patients are fully covered under both 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 $\tilde{\lambda}(\mu^o) \leq \Lambda \leq \tilde{\lambda}(\tilde{r}_f)$[20], then the optimal service rates under the FFS and BP schemes are equal (i.e., $\tilde{\mu}_f = \tilde{\mu}_b = \max\{\mu : \tilde{\lambda}(\mu) \geq \Lambda\}$). Consequently, the patient welfare, the readmission rate, the initial admission rate, the waiting time per visit and the total waiting time are the same under both FFS and BP schemes.*

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

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 patient welfare, service quality and the congestion level. These results can be explained as follows. When $\tilde{\lambda}(\mu^o) \leq \Lambda \leq \tilde{\lambda}(\tilde{r}_f)$, Proposition 4 reveals that the HCP will select the largest service rate that makes the potential patients fully covered under both 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.

[20] Note that the existence of $\tilde{\lambda}(\mu^o) \leq \Lambda \leq \tilde{\lambda}(\tilde{r}_f)$ implicitly requires that $\tilde{\lambda}(\mu^o) \leq \tilde{\lambda}(\tilde{r}_f)$. We have shown in Appendix A that there exist thresholds $\bar{t}$ and $\bar{B}$ (where the expression of $\bar{t}$ and $\bar{B}$ can be found in (27) in Appendix A) such that $\tilde{\lambda}(\mu^o) \leq \tilde{\lambda}(\tilde{r}_f)$ if and only if $t \geq \bar{t}$ and $B \geq \bar{B}$.

Next, when $\Lambda < \min\{\tilde{\lambda}(\mu^o), \tilde{\lambda}(\tilde{r}_f)\}$, μ^o is feasible under the full coverage and thus, $\tilde{\mu}_b = \mu^o$. 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 readmission 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, $T(\tilde{\mu}_b) < T(\tilde{\mu}_f)$. Thus, compared with the FFS scheme, the service quality is better and the total waiting cost is smaller under the BP scheme. Therefore, the patient welfare under the BP scheme is also 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 schemes in terms of the patient welfare, service quality and the total waiting time.

Remark. *When $\tilde{\lambda}(\tilde{r}_f) < \Lambda \leq \tilde{\lambda}(\tilde{r}_b)$ so that the potential patients are fully covered under the BP scheme but are partially covered under the FFS scheme, the comparison results are similar to Proposition 3 and Corollary 4.*[21] *Specifically, when $\tilde{\lambda}(\mu^o) \leq \tilde{\lambda}(\tilde{r}_f)$, the results given in Proposition 3 still hold: the BP scheme dominates the FFS scheme in terms of the patient 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 3 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 patient welfare, service quality and the total waiting time).*

6 Conclusion Remarks

In this paper, we have presented a queueing model with endogenous arrival rate selected by the patient and endogenous readmission rate controlled by the HCP (via the selected service rate). By analysing 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 trade-off between service rate and service quality (in terms of the readmission rate), we obtain the following managerial insights. First, when potential patients are partially covered, we find that a higher service rate may reduce both the initial admission rate and the effective admission 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

[21] See Proposition B7 in the online Appendix B for details.

readmission rate but it can increase the waiting time under both 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 patient welfare and service quality (i.e., the readmission 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 the patient 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 readmission rate, the waiting time per visit, the total waiting time and the patient welfare. Overall, the implications of our findings are as follows. First, when the size of the patient population is large, shifting from FFS to BP can improve the patient welfare and reduce readmissions, 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 FFS to BP. However, when the size of the patient population is very small, the BP scheme dominates the FFS 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 readmission rate) and service speed. However, our model has several limitations that we shall leave them as future research for 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, our model is more suitable for elective non-urgent outpatient care. However, future research is needed to examine the implications of different payment schemes on urgent care especially when the patient's health condition may deteriorate rapidly over time so that longer waiting time may adversely affect patient's health quality outcomes.

Another important area is to consider the competition among different HCPs. In the presence of the market competition, the non-cured patients may seek admissions from other HCPs. (For example, Andritsos and Tang (2014) examine the impact of competition between the private and public HCPs on the patient welfare in Europe.) Therefore, the HCP may not generate more demand by reducing readmission rate. In view of this, it is of interest to investigate the impact of competition on the HCPs' choices of service rate and other performance metrics such as the waiting time, the patient welfare and the admission rate. Finally, another possible extension is to add the mortality rate into our analysis. An implicit assumption in our model is that patients can definitely be cured in the long run. This assumption is realistic for the elective surgeries. However, for serious illness such as diabetes mellitus, the increase in the mortality rate is an inevitable consequence of the low service quality (i.e., high readmission rate and/or long waiting time).

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. Mamaniy, S. Nassiri. 2014. Bundled payment vs. Fee-for-Service: impact of payment scheme on performance. Forthcoming in *Management Science.*

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.

Bavafa, H., S. Savin, C. Terwiesch. 2013. Managing office revisit 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.*

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.

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

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.

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.

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

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.

Kociol, R. D., R. D. Lopes, R. Clare, L. Thomas, R. H. Mehta, P. Kaul, et al. 2012. International variation in and factors associated with hospital readmission after myocardial infarction. *The Journal of the American Medical Association* **307**(1) 66-74.

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. 2015. Speed-quality competition in customer-intensive services with boundedly rational customer. Forthcoming in *Production and Operations Management* .

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.

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.

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. Medicares 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.

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.

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 Appendices

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

Appendix A: Proofs of Lemmas and Propositions

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 quasi-concave 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,

$$\frac{dg(\mu)}{d\mu} = \frac{\delta^{''}(\mu)(1-\delta(\mu)) + (\delta^{'}(\mu))^2}{(1-\delta(\mu))^2} > 0, \tag{18}$$

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 $\mu_1 = \max\{\tilde{\lambda}(\mu)\}$ and $\mu_2 = \max\{\tilde{\lambda}_e(\mu)\}$. By noting that $o(\mu) = \mu(1-\delta(\mu))$ and $o^{'}(\mu) = 1-\delta(\mu) - \mu\delta^{'}(\mu)$, we have $o(\mu) - \mu o^{'}(\mu) = \mu^2\delta^{'}(\mu)$. Hence, taking derivative of $\tilde{\lambda}(\mu)$ with respect to μ yields

$$\begin{aligned}\frac{d\tilde{\lambda}(\mu)}{d\mu} &= o^{'}(\mu) - \frac{\theta o^{'}(\mu)}{Ro(\mu) - \mu t} + \frac{\theta o(\mu)(Ro^{'}(\mu) - t)}{[Ro(\mu) - \mu t]^2}\\ &= o^{'}(\mu) - \frac{\theta\delta^{'}(\mu)t}{[R(1-\delta(\mu)) - t]^2}.\end{aligned} \tag{19}$$

From Lemma 1, we can show that $d\tilde{\lambda}(\mu)/d\mu < 0$ for $\mu \geq \mu^o$. Thus $\mu_1 < \mu^o$. Using Lemma 1 again, $o^{''}(\mu_1) < 0$. Recall from (18) that $\delta^{''}(\mu)(1-\delta(\mu)) + (\delta^{'}(\mu))^2 > 0$. Therefore,

$$\begin{aligned}
\left.\frac{d^2\tilde{\lambda}(\mu)}{d\mu^2}\right|_{\mu=\mu_1} &= o^{''}(\mu_1) - \frac{\theta t\delta^{''}(\mu_1)}{[R(1-\delta(\mu_1))-t]^2} - \frac{2R\theta(\delta^{'}(\mu_1))^2 t}{[R(1-\delta(\mu_1))-t]^3} \\
&= o^{''}(\mu_1) + \frac{\theta t(\delta^{'}(\mu_1))^2}{(1-\delta(\mu_1))[R(1-\delta(\mu_1))-t]^2} - \frac{2R\theta(\delta^{'}(\mu_1))^2 t}{[R(1-\delta(\mu_1))-t]^3} \\
&\quad -\frac{\theta t[\delta^{''}(\mu_1)(1-\delta(\mu_1)) + (\delta^{'}(\mu_1))^2]}{(1-\delta(\mu_1))[R(1-\delta(\mu_1))-t]^2} \\
&= o^{''}(\mu_1) - \frac{\theta t[\delta^{''}(\mu_1)(1-\delta(\mu_1)) + (\delta^{'}(\mu_1))^2]}{(1-\delta(\mu_1))[R(1-\delta(\mu_1))-t]^2} - \frac{\theta t(\delta^{'}(\mu_1))^2(R(1-\delta(\mu_1))+t)}{(1-\delta(\mu_1))[R(1-\delta(\mu_1))-t]^3} \qquad (20) \\
&< 0,
\end{aligned}$$

which shows that $\tilde{\lambda}(\mu)$ is quasi-concave in μ. Taking derivative of $\tilde{\lambda}_e(\mu)$ with respect to μ and using (18),

$$\begin{aligned}
\frac{d\tilde{\lambda}_e(\mu)}{d\mu} &= 1 - \frac{\theta R\delta^{'}(\mu)}{[R(1-\delta(\mu))-t]^2}, \qquad (21) \\
\frac{d^2\tilde{\lambda}_e(\mu)}{d\mu^2} &= -\frac{\theta R\delta^{''}(\mu)}{[R(1-\delta(\mu))-t]^2} - \frac{2\theta R^2(\delta^{'}(\mu))^2}{[R(1-\delta(\mu))-t]^3} \\
&= \frac{-R\theta}{(1-\delta(\mu))}\left[\frac{(\delta^{'}(\mu))^2(R(1-\delta(\mu))+t)}{[R(1-\delta(\mu))-t]^3} + \frac{\delta^{''}(\mu)(1-\delta(\mu)) + (\delta^{'}(\mu))^2}{[R(1-\delta(\mu))-t]^2}\right] \\
&< 0. \qquad (22)
\end{aligned}$$

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, $\mu_1 < \mu_2$. By noting that $o^{'}(\mu) = 1 - \delta(\mu) - \mu\delta^{'}(\mu)$, (21) can be rewritten as

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

Obviously, the maximum effective admission rate should be positive; that is, $\tilde{\lambda}_e(\mu_2) > 0$. Because μ_2 should satisfy the FOC of $\tilde{\lambda}_e(\mu,)$, namely, $d\tilde{\lambda}_e(\mu_2)/d\mu = 0$, we have $Ro^{'}(\mu_2) - t < 0$. Using (21) and substituting $d\tilde{\lambda}_e(\mu_2)/d\mu = 0$ into (19), we have

$$\left.\frac{d\lambda(\mu)}{d\mu}\right|_{\mu=\mu_2} = \frac{Ro^{'}(\mu_2)-t}{R} < 0, \qquad (23)$$

which implies that $\mu_2 > \mu_1$. □

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

$$\frac{d\Pi_f(\mu)}{d\mu} = \frac{c\tilde{\lambda}_e(\mu)}{\mu^2} + \left(r_f - \frac{c}{\mu}\right)\frac{d\tilde{\lambda}_e(\mu)}{d\mu} = 0, \tag{24}$$

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,

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

which shows that $\Pi_f(\mu)$ is quasi-concave 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

$$\frac{d\Pi_b(\mu)}{d\mu} = \frac{co^{'}(\mu)\tilde{\lambda}(\mu)}{(o(\mu))^2} + \left(r_b - \frac{c}{o(\mu)}\right)\frac{d\tilde{\lambda}(\mu)}{d\mu}. \tag{25}$$

From (19) 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, then we have $o^{'}(\tilde{\mu}_b) > 0$ and $o^{''}(\tilde{\mu}_b) < 0$. Then from (20) and (25), 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 (19), 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 notion, we interchangeably use $\tilde{\mu}_s(r_s)$ and $\tilde{\mu}_s$. Differentiating (24) 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 theory,

$$\frac{d\tilde{\mu}_f}{dr_f} = -\left.\frac{\frac{\partial^2\Pi_f(\mu)}{\partial\mu\partial r_f}}{\frac{\partial^2\Pi_f(\mu)}{\partial\mu^2}}\right|_{\mu=\tilde{\mu}_f} < 0.$$

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 (25) 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 theory again, we have

$$\frac{d\tilde{\mu}_b}{dr_b} = -\left.\frac{\frac{\partial^2\Pi_b(\mu)}{\partial\mu\partial r_b}}{\frac{\partial^2\Pi_b(\mu)}{\partial\mu^2}}\right|_{\mu=\tilde{\mu}_b} < 0.$$

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

$$W(r_s) = \frac{R(1-\delta(\tilde{\mu}_s))}{\theta} - \frac{t}{\theta};\ T(r_s) \quad = \quad \frac{R}{\theta} - \frac{t}{\theta(1-\delta(\tilde{\mu}_s))}.$$

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, $\tilde{\lambda}(r_f)$ and $r_f\tilde{\lambda}_e(r_f)$ are increasing in r_f, and $\tilde{\lambda}(r_b)$ and $r_b\tilde{\lambda}(r_b)$ are increasing in r_b. From (14), the funder's objective is equivalent to maximizing the initial admission rate $\tilde{\lambda}(r_s)$. Therefore, the budget constraint (15) will be binding and the funder's optimal reimbursement rate satisfies $\tilde{r}_f\tilde{\lambda}_e(\tilde{r}_f) = B$ under FFS scheme and $\tilde{r}_b\tilde{\lambda}(\tilde{r}_b) = B$ under BP scheme. □

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 admission 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. We 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 Propositions 2, under the partial coverage, the budget constraints associated with both 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

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

since $\tilde{\lambda}(\tilde{r}_f) \geq \tilde{\lambda}(\tilde{r}_b)$. 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

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

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 (14), 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

$$W(\mu) = \frac{R(1-\delta(\mu))}{\theta} - \frac{t}{\theta};\ T(\mu) \quad = \quad \frac{R}{\theta} - \frac{t}{\theta(1-\delta(\mu))}, \tag{26}$$

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

Proof of Proposition 4. Differentiating $\Pi_f(\mu)$ given in (16) 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 (17) 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 Corollary 4. According to Proposition 4, when $\tilde{\lambda}(\mu^o) \leq \tilde{\lambda}(\tilde{r}_f)$ and $\tilde{\lambda}(\mu^o) \leq \Lambda \leq \tilde{\lambda}(\tilde{r}_f)$, $\tilde{\mu}_f = \tilde{\mu}_b = \max\{\mu|\tilde{\lambda}(\mu) \geq \Lambda\}$. While when $\Lambda < \min\{\tilde{\lambda}(\mu^o), \tilde{\lambda}(\tilde{r}_f)\}$, $\tilde{\mu}_b = \mu^o$ and $\tilde{\mu}_f = \max\{\mu|\tilde{\lambda}(\mu) \geq \Lambda\}$. Since $\max\{\mu|\tilde{\lambda}(\mu) \geq \Lambda\}$ 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, $T(\Lambda, \tilde{\mu}_b) < T(\Lambda, \tilde{\mu}_f)$. Also, since $\delta(\mu)$ is increasing in μ, $\delta(\tilde{\mu}_f) > \delta(\tilde{\mu}_b)$.

In addition, 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 about the patient's utility:

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

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

Below we will show that there exist thresholds $\bar{t}$ and $\bar{B}$ such that under the partial coverage scenario, $\tilde{\lambda}(\tilde{r}_f) \geq \tilde{\lambda}(\mu^o)$ if and only if (iff) $t \geq \bar{t}$ and $B \geq \bar{B}$. To this end, we first show that when $\frac{d\tilde{\lambda}_e(\mu)}{d\mu}\big|_{\mu=\mu^o} \geq 0$, $\tilde{\lambda}(\tilde{r}_f) < \tilde{\lambda}(\mu^o)$. We have shown in the proof of Proposition 1 that $\frac{d\tilde{\lambda}_e(\mu)}{d\mu}\big|_{\mu=\tilde{\mu}_f(\tilde{r}_f)} < 0$. Because $\tilde{\lambda}_e(\mu)$ is unimodal in μ,this implies that $\mu^o < \tilde{\mu}_f(\tilde{r}_f)$. Also, from the proof of Proposition 3 we have $\frac{d\tilde{\lambda}(\tilde{\mu}_f(\tilde{r}_f))}{d\mu} < 0$. In addition, from (19), we get $\frac{d\tilde{\lambda}(\mu^o)}{d\mu} < 0$ since $o^{'}(\mu^o) = 0$. As $\tilde{\lambda}(\mu)$ is unimodal in μ and $\mu^o < \tilde{\mu}_f(\tilde{r}_f)$, we have $\tilde{\lambda}(\tilde{r}_f) < \tilde{\lambda}(\mu^o)$. Thus, $\tilde{\lambda}(\tilde{r}_f) \geq \tilde{\lambda}(\mu^o)$ requires that $\frac{d\tilde{\lambda}_e(\mu)}{d\mu}\big|_{\mu=\mu^o} < 0$.

Next, we show that $\frac{d\tilde{\lambda}_e(\mu)}{d\mu}\big|_{\mu=\mu^o} < 0$ iff $t > R \cdot (1 - \delta(\mu^o)) - \sqrt{\theta R \delta'(\mu^o)}$. Substituting $t = R \cdot (1 - \delta(\mu^o)) - \sqrt{\theta R \delta'(\mu^o)}$ into (21), we have $\frac{d\tilde{\lambda}_e(\mu)}{d\mu}\big|_{\mu=\mu^o} = 0$. From (21),

$$\frac{\partial^2 \tilde{\lambda}_e(\mu)}{\partial \mu \partial t} = -\frac{2\theta R \delta^{'}(\mu)}{[R(1 - \delta(\mu)) - t]^3} < 0.$$

Thus, $\frac{d\tilde{\lambda}_e(\mu)}{d\mu}\big|_{\mu=\mu^o} < 0$ iff $t > R \cdot (1 - \delta(\mu^o)) - \sqrt{\theta R \delta'(\mu^o)}$.

Finally, we show that when $\frac{d\tilde{\lambda}_e(\mu)}{d\mu}\big|_{\mu=\mu^o} < 0$, there exists a $\bar{B}$ such that $\tilde{\lambda}(\tilde{r}_f) \geq \tilde{\lambda}(\mu^o)$ iff $B \geq \bar{B}$. To this end, we first show that there exists a $r_f^{'}$ such that $\tilde{\mu}_f(r_f^{'}) = \mu^o$. First, we can show that when $r_f = c/\mu^o$, the HCP's corresponding optimal service rate should satisfy $\tilde{\mu}_f(r_f)|r_f = c/\mu^o > \mu^o$ as the HCP's marginal profit (i.e., $r_f - c/\mu$) shall be positive. Next, the FOC (24) implies that when r_f goes to infinity, $\frac{d\tilde{\lambda}_e(\tilde{\mu}_f(r_f))}{d\mu} \to 0$. As now $\frac{d\tilde{\lambda}_e(\mu)}{d\mu}\big|_{\mu=\mu^o} < 0$, $\lim_{r_f \to +\infty} \tilde{\mu}_f(r_f) < \mu^o$. Since $\tilde{\mu}_f(r_f)$ is decreasing in r_f (Corollary 2), there exists a $r_f^{'}$ such that $\tilde{\mu}_f(r_f^{'}) = \mu^o$. Recall from Proposition 2 that in equilibrium, the budget constraint under the partial coverage is binding, that is, $\tilde{r}_f \tilde{\lambda}_e(\tilde{r}_f) = B$. Because $\tilde{\lambda}(r_f)$ is increasing in r_f and $\tilde{\mu}_f(r_f^{'}) = \mu^o$, when $B \geq r_f^{'} \tilde{\lambda}_e(\mu^o)$, $\tilde{r}_f \geq r_f^{'}$. As $\tilde{\mu}_f(r_f)$ is decreasing in r_f, $\tilde{\mu}_f(\tilde{r}_f) \leq \tilde{\mu}_f(r_f^{'}) = \mu^o$. We have shown in the proof of Proposition 3 that $\frac{d\tilde{\lambda}(\tilde{\mu}_f(\tilde{r}_f))}{d\mu} < 0$. And from (19), we get $\frac{d\tilde{\lambda}(\mu^o)}{d\mu} < 0$ as $o^{'}(\mu^o) = 0$. As $\tilde{\lambda}(\mu)$ is unimodal in μ and $\tilde{\mu}_f(\tilde{r}_f) \leq \mu^o$, $\tilde{\lambda}(\tilde{r}_f) \geq \tilde{\lambda}(\mu^o)$. Similarly, when $B < r_f^{'} \tilde{\lambda}_e(\mu^o)$, we can show that $\tilde{r}_f < r_f^{'}$, $\tilde{\mu}_f(\tilde{r}_f) > \tilde{\mu}_f(r_f^{'}) = \mu^o$ and $\tilde{\lambda}(\tilde{r}_f) < \tilde{\lambda}(\mu^o)$. Let

$$\bar{t} = R \cdot (1 - \delta(\mu^o)) - \sqrt{\theta R \delta'(\mu^o)} \text{ and } \bar{B} \;=\; r_f^{'} \tilde{\lambda}_e(\mu^o), \text{ where } \tilde{\mu}_f(r_f^{'}) = \mu^o. \tag{27}$$

Then $\tilde{\lambda}(\tilde{r}_f) \geq \tilde{\lambda}(\mu^o)$ iff $t \geq \bar{t}$ and $B \geq \bar{B}$. □

Appendix B: The Analysis of the Scenario When the Potential Arrival Rate is Small

In this appendix, we study the scenario under which the potential patient size is relatively small such that it is possible that no patients will leave without being treated. Again we apply the backward induction to derive the patients' queueing-joining decisions, the HCP's service rate decisions and the funder's reimbursement decisions under both the FFS and BP schemes.

B.1: Endogenous Patient Initial Admissions

Recall that if the potential initial admission rate is relatively large such that $U(\Lambda, \mu) < 0$, the potential patients are partially covered and the health care system reaches the partial coverage scenario. While if the potential initial admission rate is relatively small such that $U(\Lambda, \mu) \geq 0$, all the patients choose to seek admissions from the HCP and the health care system reaches the full coverage scenario. We have derived the initial admission rate $\tilde{\lambda}(\mu)$ and the effective admission rate $\tilde{\lambda}_e(\mu)$ under the partial coverage case in the main context of this paper. In this appendix, we still use $\tilde{\lambda}(\mu)$ and $\tilde{\lambda}_e(\mu)$ to represent the initial admission rate and the effective admission rate under the partial coverage scenario, respectively. However, to avoid confusion, hereafter we let $\hat{\lambda}(\mu)$ and $\hat{\lambda}_e(\mu)$ represent the initial admission rate and the effective admission rate in equilibrium, respectively. Taking into account that the full coverage case may occur in the equilibrium, based on (7), (8) and (9), we then have

$$\hat{\lambda}(\mu) = \min\{\Lambda, \tilde{\lambda}(\mu)\} \text{ and } \hat{\lambda}_e(\mu) = \min\{n(\mu)\Lambda, \tilde{\lambda}_e(\mu)\}. \tag{28}$$

Next, we will use (28) to derive the HCP's service rate decision and the funder's reimbursement decision. To facilitate our analysis, we shall analyze both the partial coverage scenario and the full coverage scenario separately. Since we have analyzed the partial coverage scenario in the main context, here we will focus on analyzing the full coverage scenario. To be consistent, we still use " $\tilde{\cdot}$ " to indicate the variables associated with the partial coverage scenario and use " $\hat{\cdot}$ " to indicate the equilibrium outcomes.

B.2 The FFS Scheme

We first consider the FFS scheme. In this section, we first determine the HCP's optimal service rate $\hat{\mu}_f$. Then, by anticipating the HCP's service rate decision and the patient's admission behavior, we determine the funder's reimbursement decision $\hat{r}_f$.

B.2.1 The HCP's Service Rate Decision

From (28), we know that the potential patients are fully covered iff $\tilde{\lambda}(\mu) \geq \Lambda$. When the patients are fully covered, $\hat{\lambda}(\mu) = \Lambda$ and $\hat{\lambda}_e(\mu) = \Lambda/(1 - \delta(\mu))$. By substituting $\hat{\lambda}(\mu) = \Lambda$ and $\hat{\lambda}_e(\mu) = \Lambda/(1 - \delta(\mu))$ into (10) and taking the full coverage requirement $\tilde{\lambda}(\mu) \geq \Lambda$ into account, we can write the the HCP's optimization problem under the full coverage case as

$$\max_{\mu} \Pi_f(\mu) = \left(r_f - \frac{c}{\mu}\right)\frac{\Lambda}{1-\delta(\mu)}, \tag{29}$$

$$s.t. \qquad \tilde{\lambda}(\mu) \geq \Lambda. \tag{30}$$

A close look of (29) reveals that as the service rate μ increases, the variable cost c/μ decreases while the demand $\Lambda/(1-\delta(\mu))$ increases. Therefore, the HCP's profit $\Pi_f(\mu)$ is increasing in μ. Because $\tilde{\lambda}(\mu)$ is unimodal in μ (according to Corollary 1), there exist at most two service rates such that the constraint (30) is binding (i.e., $\tilde{\lambda}(\mu) = \Lambda$). And $\tilde{\lambda}(\mu) \geq \Lambda$ iff μ is located between them. As the HCP's profit $\Pi_f(\mu)$ increases in μ, the optimal service rate under the full coverage shall be the larger root of $\tilde{\lambda}(\mu) = \Lambda$. In other words, the HCP chooses the largest service rate that satisfies the full coverage requirement.

Proposition B1. *Suppose the potential patients are fully covered such that $\hat{\lambda}(\mu) = \Lambda$. Under the FFS scheme, for a given reimbursement rate r_f, the HCP's profit $\Pi_f(\mu)$ given in* (29) *is increasing in μ. The optimal service rate equals $\bar{\mu}$, where $\bar{\mu}$ is the larger root of $\tilde{\lambda}(\mu) = \Lambda$ and is independent of r_f.*

B.2.1.1 Equilibrium Service Rate

So far we have derived the HCP's optimal service rate and the corresponding expected profit under both partial and full coverage scenarios. By comparing the HCP's expected profit under these two scenarios, we can derive the HCP's optimal service rate under the FFS scheme as follows.

Proposition B2. *Under the FFS scheme, for a given reimbursement rate r_f, the HCP's optimal service rate satisfies*

$$\hat{\mu}_f(r_f) = \begin{cases} \tilde{\mu}_f(r_f), & \text{if } \tilde{\lambda}(r_f) < \Lambda, \\ \bar{\mu}, & \text{if } \tilde{\lambda}(r_f) \geq \Lambda. \end{cases} \tag{31}$$

Proposition B2 reveals that if the potential patient size is relatively large (i.e., $\tilde{\lambda}(r_f) < \Lambda$), then some patients will balk and the system reaches the partial coverage. While if the potential patient size is relatively small (i.e., $\tilde{\lambda}(r_f) \geq \Lambda$), then the system ends up being fully covered.

B.2.2 The Funder's Reimbursement Decision

Anticipating the HCP's service rate $\hat{\mu}_f(r_f)$ given in (31), we now turn our attention to determining the funder's reimbursement decision. Similarly, we first discuss the partial coverage scenario and the full coverage scenario separately. Then, by comparing the patient welfare associated with these two scenarios, we can figure out under which conditions which scenario appears as the equilibrium outcome. Since we have analyzed the partial coverage scenario in the main context, we now focus on the full coverage scenario. Plugging $\tilde{\lambda}(\mu) = \Lambda$ and $\bar{\mu}$ into (14), the funder's problem under the full coverage scenario is

$$\begin{aligned} \max_{r_f} S(r_f) &= \Lambda \cdot U(\Lambda) = \Lambda \cdot (R - \theta \cdot T(\Lambda, \bar{\mu}) - t \cdot n(\bar{\mu})), && (32) \\ s.t. \qquad & r_f \cdot \frac{\Lambda}{1-\delta(\bar{\mu})} \leq B, && (33) \end{aligned}$$

Recall from Proposition B1 that the HCP's optimal service rate $\bar{\mu}$ under the full coverage is independent of r_f. Therefore, the funder fails to regulate the HCP's service rate decision when the potential patients are fully covered. However, the threshold $\tilde{\lambda}(r_f)$ (in (31)) which determines when the full coverage scenario occurs does depend on r_f. Thus, the funder can indirectly control the full coverage scenario by regulating the threshold $\tilde{\lambda}(r_f)$.

B.2.3 Equilibrium Outcome

So far we have derived the optimal equilibrium outcomes associated with the partial coverage scenario and the full coverage scenario, respectively. Below we further investigate under which conditions which scenario appears as the equilibrium outcome.

Proposition B3. *The equilibrium outcome associated with the FFS scheme can be characterized as follows.*

1. *When $\tilde{\lambda}(\tilde{r}_f) < \Lambda$, the potential patients under the FFS scheme in equilibrium are partially covered. Therefore, in equilibrium, the funder's reimbursement rate $\hat{r}_f$ satisfies $\hat{r}_f = \tilde{r}_f$ and the HCP's service rate $\hat{\mu}_f(\hat{r}_f)$ satisfies $\hat{\mu}_f(\hat{r}_f) = \tilde{\mu}_f(\tilde{r}_f)$.*

2. *When $\tilde{\lambda}(\tilde{r}_f) \geq \Lambda$, the potential patients under the FFS scheme in equilibrium are fully covered. Thus, in equilibrium, the funder's reimbursement rate $\hat{r}_f$ satisfies $\tilde{\mu}_f(\hat{r}_f) = \bar{\mu}$ and the HCP's service rate $\hat{\mu}_f(\hat{r}_f) = \bar{\mu}$.*

Proposition B3 reveals that when the potential initial arrival rate Λ is large enough such that $\tilde{\lambda}(\tilde{r}_f) < \Lambda$, some patients in equilibrium will leave without being treated and the health care system will end up with the partial coverage (i.e., $\hat{r}_f = \tilde{r}_f$ and $\hat{\mu}_f(\hat{r}_f) = \tilde{\mu}_f(\tilde{r}_f)$). While if the potential initial arrival rate Λ is relatively small such that $\tilde{\lambda}(\tilde{r}_f) \geq \Lambda$, all the patients can receive the treatment and the health care system will end up with the full coverage. Recall from Proposition B1 that the HCP's optimal service rate under the full coverage $\bar{\mu}$ satisfies $\tilde{\lambda}(\bar{\mu}) = \Lambda$ and is independent of r_f. The patient welfare under the full coverage remains the same for any feasible reimbursement rate r_f. However, the funder will not be more generous than it has to be. Therefore, it will choose the smallest feasible reimbursement rate under the full coverage. Since $\tilde{\lambda}(r_f)$ is increasing in r_f (according to Corollary 3), the funder will select the reimbursement rate that just satisfies the full coverage requirement $\tilde{\lambda}(\tilde{\mu}_f(\hat{r}_f)) = \tilde{\lambda}(\bar{\mu}) = \Lambda$.

B.3 The BP Scheme

In this section, we study the BP scheme. Similar to §B.2, to solve the equilibrium outcome associated with the BP scheme, we first determine the HCP's optimal service rate $\hat{\mu}_b$. Then by anticipating the HCP's service rate decision and the patient's admission behavior, we determine the funder's reimbursement decision $\hat{r}_b$. Also, we will consider the partial coverage scenario and the full coverage scenario separately. Since the partial coverage scenario has been analyzed in the main context, we will focus on the full coverage scenario in the following.

B.3.1 The HCP's Service Rate Decision

From (28), we know that to ensure the full coverage of potential patients, the HCP's service rate should satisfy $\tilde{\lambda}(\mu) \geq \Lambda$. Plugging $\hat{\lambda}(\mu) = \Lambda$ and $\hat{\lambda}_e(\mu) = \Lambda/(1-\delta(\mu))$ into (11) and taking the full coverage requirement $\tilde{\lambda}(\mu) \geq \Lambda$ into consideration, the HCP's problem under the full coverage becomes

$$\max_{\mu} \Pi_b(\mu) \quad = \quad \max_{\mu} \left(r_b - \frac{c}{o(\mu)} \right) \Lambda, \tag{34}$$

$$s.t. \qquad \tilde{\lambda}(\mu) \geq \Lambda. \tag{35}$$

In (34), the HCP's objective under the full coverage 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 $\mu = \mu^o$. Therefore, $\Pi_b(\mu)$ given in (34) is also unimodal in μ and reaches its maximum at $\mu = \mu^o$. When the size of potential patients is small (i.e., $\tilde{\lambda}(\mu^o) \geq \Lambda$), the service rate μ^o that maximizes $\Pi_b(\mu)$ given in (34) is feasible under the full coverage. Thus, the HCP will set $\mu = \mu^o$ in equilibrium. Whereas, when the potential initial admission rate is large (i.e., $\tilde{\lambda}(\mu^o) < \Lambda$), μ^o is infeasible and therefore, the full coverage requirement $\tilde{\lambda}(\mu) \geq \Lambda$ will be binding. Since $\tilde{\lambda}(\mu)$ is unimodal in μ, there exist two roots of $\tilde{\lambda}(\mu) = \Lambda$. As one can easily show that the second term of $\tilde{\lambda}(\mu)$ in (8) (i.e., $\theta(1-\delta(\mu))/R(1-\delta(\mu)) - t$) is increasing in μ, $o(\mu)$ associated with the larger root of $\tilde{\lambda}(\mu) = \Lambda$ shall be larger. Recall that under the full coverage scenario, the HCP's objective associated with the BP scheme is equivalent to minimizing the variable cost $c/o(\mu)$. The HCP then shall choose the larger root of $\tilde{\lambda}(\mu) = \Lambda$. Thus, we obtain the following results.

Proposition B4. *Suppose the potential patients are fully covered such that $\hat{\lambda}(\mu) = \Lambda$. Under the BP scheme, for a given reimbursement rate r_b, the HCP's profit $\Pi_b(\mu)$ given in* (34) *is unimodal in μ. Furthermore,*

$$\text{the HCP's optimal service rate} = \begin{cases} \bar{\mu}, & \text{if } \tilde{\lambda}(\mu^o) < \Lambda, \\ \mu^o, & \text{if } \tilde{\lambda}(\mu^o) \geq \Lambda, \end{cases} \tag{36}$$

where $\bar{\mu}$ is the larger root of $\tilde{\lambda}(\mu) = \Lambda$ and is independent of r_b.

B.3.1.2 Equilibrium Service Rate

In this section, we will derive the HCP's optimal service rate under the BP scheme by comparing the HCP's expected profit associated with the partial and full coverage scenarios, respectively. When the potential initial admission rate is very large (i.e., $\tilde{\lambda}(r_b) < \Lambda$), the optimal service rate $\tilde{\mu}_b(r_b)$ under the partial coverage is feasible and it is the equilibrium service rate. When the potential initial admission rate is relatively small such that the optimal service rate under the partial coverage is infeasible (i.e., $\tilde{\lambda}(r_b) \geq \Lambda$), the system ends up with the full coverage. In this case, by noting from Proposition 1 that $\tilde{\lambda}(\mu^o) < \tilde{\lambda}(r_b)$, then if the potential initial admission rate is small enough (i.e., $\tilde{\lambda}(\mu^o) > \Lambda$), the HCP will select the service rate μ^o to minimize its variable cost. But if the potential initial admission rate is in a moderate range (i.e, $\tilde{\lambda}(\mu^o) \leq \Lambda \leq \tilde{\lambda}(r_b)$), the full coverage requirement will be binding (i.e., $\tilde{\lambda}(\tilde{\mu}_b(r_b)) = \tilde{\lambda}(\bar{\mu}) = \Lambda$). We then have the following results.

Proposition B5. *Under the BP scheme, for a given reimbursement rate r_b, the HCP's optimal service rate satisfies*

$$\hat{\mu}_b(r_b) = \begin{cases} \tilde{\mu}_b(r_b), & \text{if } \tilde{\lambda}(r_b) < \Lambda, \\ \bar{\mu}, & \text{if } \tilde{\lambda}(\mu^o) < \Lambda \leq \tilde{\lambda}(r_b), \\ \mu^o, & \text{if } \tilde{\lambda}(\mu^o) \geq \Lambda. \end{cases} \tag{37}$$

B.3.2 The Funder's Reimbursement Decision

We now turn our attention to determining the funder's reimbursement decision under the BP scheme. Anticipating the HCP's service rate $\tilde{\mu}_b(r_b)$ given in (37), the funder selects the reimbursement rate r_b to maximize the patient welfare $S(r_b)$ in equilibrium. Plugging $\tilde{\lambda}(\mu) = \Lambda$ and the HCP's optimal service rate given in (36) into (14), the funder's problem under the full coverage becomes

$$\max_{r_b} S(r_b) = \Lambda \cdot U(\Lambda) = \begin{cases} \Lambda \cdot (R - \theta \cdot T(\Lambda, \bar{\mu}) - t \cdot n(\bar{\mu})), & \text{if } \tilde{\lambda}(\mu^o) < \Lambda, \\ \Lambda \cdot (R - \theta \cdot T(\Lambda, \mu^o) - t \cdot n(\mu^o)), & \text{if } \tilde{\lambda}(\mu^o) \geq \Lambda, \end{cases} \tag{38}$$

$$s.t. \quad \begin{cases} r_b \cdot \frac{\Lambda}{1-\delta(\bar{\mu})} \leq B, & \text{if } \tilde{\lambda}(\mu^o) < \Lambda, \\ r_b \cdot \frac{\Lambda}{1-\delta(\mu^o)} \leq B, & \text{if } \tilde{\lambda}(\mu^o) \geq \Lambda. \end{cases} \tag{39}$$

Because both $\bar{\mu}$ and μ^o are independent of r_b, the funder fails to regulate the HCP's service rate decisions when the potential initial patients are fully covered. However, the threshold $\tilde{\lambda}(r_b)$ in (37) which determines when the full coverage scenario occurs does depend on r_b. Therefore, the funder can control the occurrence of the full coverage scenario by regulating the threshold $\tilde{\lambda}(r_b)$.

B.3.2.1 Equilibrium Outcome

Based on the optimal decisions of the patients, the HCP, and the funder associated with the partial and full coverage scenarios, we now investigate under which conditions which scenario appears as the equilibrium outcome.

Proposition B6. *The equilibrium outcome associated with the BP scheme can be characterized as follows.*

1. *When $\tilde{\lambda}(\tilde{r}_b) < \Lambda$, the potential patients under the BP scheme in equilibrium are partially covered . Therefore, in equilibrium, the funder's reimbursement rate $\hat{r}_b$ satisfies $\hat{r}_b = \tilde{r}_b$ and the HCP's service rate $\hat{\mu}_b(\hat{r}_b)$ satisfies $\hat{\mu}_b(\hat{r}_b) = \tilde{\mu}_b(\tilde{r}_b)$.*

2. *When $\tilde{\lambda}(\mu^o) < \Lambda \leq \tilde{\lambda}(r_b)$, the potential patients under the BP scheme in equilibrium are fully covered. In equilibrium, the funder's reimbursement rate $\hat{r}_b$ satisfies $\tilde{\mu}_b(\hat{r}_b) = \bar{\mu}$ and the HCP's service rate $\hat{\mu}_b(\hat{r}_b) = \bar{\mu}$.*

3. *When $\Lambda \leq \tilde{\lambda}(\mu^o)$, the potential patients under the BP scheme in equilibrium are also fully covered. In equilibrium, the funder's reimbursement rate $\hat{r}_b$ satisfies $\tilde{\mu}_b(\hat{r}_b) = \mu^o$ and the HCP's service rate $\hat{\mu}_b(\hat{r}_b) = \mu^o$.*

Proposition B6 shows that when the results under the partial coverage is feasible (i.e., $\tilde{\lambda}(\tilde{r}_b) < \Lambda$), the health care system will end up with the partial coverage (i.e., $\hat{r}_b = \tilde{r}_b$ and $\hat{\mu}_b(\hat{r}_b) = \tilde{\mu}_b(\tilde{r}_b)$). Otherwise, all the potential initial patients in equilibrium will seek admissions from the HCP. Recall that the patient welfare under the full coverage (given in (38)) is independent of r_b. Note that the funder will not be more generous than it has to be. Thus, under the full coverage it will choose the smallest feasible reimbursement rate. Since $\tilde{\lambda}(\bar{\mu}) = \Lambda$ and $\tilde{\lambda}(r_b)$ is increasing in r_b, $\hat{r}_b$ that satisfies $\tilde{\mu}_b(\hat{r}_b) = \bar{\mu}$ is the smallest feasible reimbursement rate under the full coverage.

B.4 Performance Comparison: FFS vs BP

So far, we have obtained the equilibrium outcomes of different performance metrics associated with the FFS and BP schemes, respectively. Because both the FFS and BP schemes may achieve either the partial or the full coverage scenario, to compare the equilibrium outcomes under these two schemes, we shall first figure out the conditions under which which scenario may occur. The corresponding conditions are characterized in the following lemma.

Lemma B1. *1. When $\Lambda > \tilde{\lambda}(\tilde{r}_b)$, the potential patients are partially covered under both FFS and BP schemes. Also, the initial admission rate associated with the BP scheme is higher than that of the FFS scheme; i.e., $\tilde{\lambda}(\tilde{r}_b) > \tilde{\lambda}(\tilde{r}_f)$.*

2. 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.

3. When $\Lambda \leq \tilde{\lambda}(\tilde{r}_f)$, the potential patients are fully covered under both FFS and BP schemes so that the endogenous initial arrival rate equals Λ.

Lemma B1 implies that to compare the performance of the two schemes, we should consider three cases: 1) when $\Lambda > \tilde{\lambda}(\tilde{r}_b)$, the potential patients are partially covered under both FFS and BP schemes; 2) when $\Lambda < \tilde{\lambda}(\tilde{r}_f)$, the potential patients are fully covered under both FFS and BP schemes; 3) 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 have examined the first two cases in the main body of this paper. In the following we will only examine the third case $\tilde{\lambda}(\tilde{r}_f) < \Lambda \leq \tilde{\lambda}(\tilde{r}_b)$.

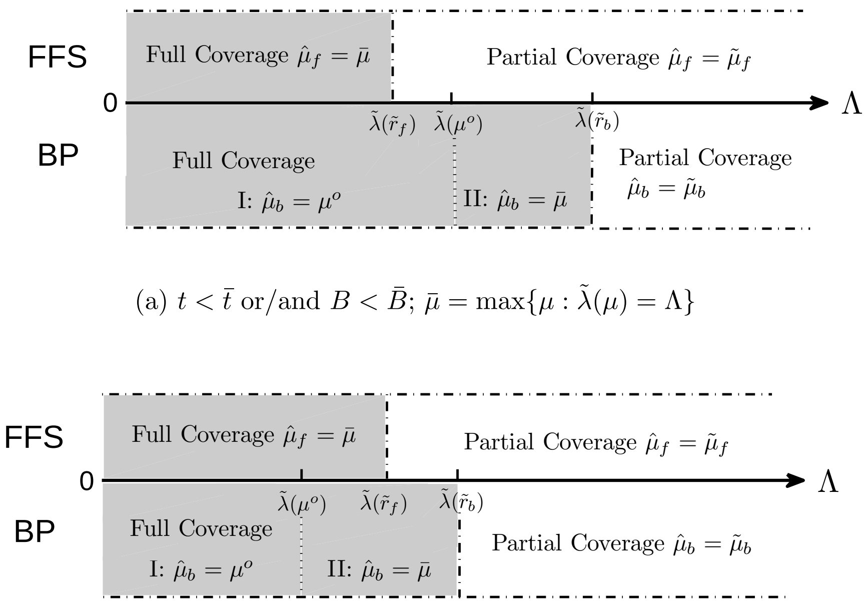
Proposition B7. *When $\tilde{\lambda}(\tilde{r}_f) < \Lambda \leq \tilde{\lambda}(\tilde{r}_b)$ such that the potential patients are fully covered under the BP scheme but are partially covered under the FFS scheme, relative to the FFS scheme,*

1. *when $\max\{\tilde{\lambda}(\tilde{r}_f), \tilde{\lambda}(\mu^o)\} \leq \Lambda \leq \tilde{\lambda}(\tilde{r}_b)$, the BP scheme dominates the FFS scheme in terms of the patient welfare (i.e., $S(\hat{r}_b) > S(\hat{r}_f)$) and service quality (i.e, the readmission rate $\delta(\hat{r}_b) < \delta(\hat{r}_f)$), while the FFS scheme outperforms the BP scheme in terms of the waiting time per visit and the total waiting time (i.e., $W(\hat{r}_b) > W(\hat{r}_f)$ and $T(\hat{r}_b) > T(\hat{r}_f)$).*

2. *if $\tilde{\lambda}(\tilde{r}_f) < \tilde{\lambda}(\mu^o)$, there exists a $\bar{\Lambda} \in [\tilde{\lambda}(\tilde{r}_f), \tilde{\lambda}(\mu^o)]$ such that*

   (a) if $\Lambda \in [\tilde{\lambda}(\tilde{r}_f), \bar{\Lambda}]$, the BP scheme dominates the FFS scheme in terms of the patient welfare, service quality and the total waiting time (i.e., $S(\hat{r}_f) < S(\hat{r}_b)$, $\delta(\hat{r}_f) > \delta(\hat{r}_b)$ and $T(\hat{r}_f) > T(\hat{r}_b)$).

   (b) if $\Lambda \in [\bar{\Lambda}, \tilde{\lambda}(\mu^o)]$, the BP scheme dominates the FFS scheme in terms of the patient welfare and service quality (i.e., $S(\hat{r}_f) < S(\hat{r}_b)$, $\delta(\hat{r}_f) > \delta(\hat{r}_b)$), while the FFS scheme outperforms the BP scheme in terms of the total waiting time (i.e., $T(\hat{r}_f) < T(\hat{r}_b)$).


FFS
Full Coverage $\hat{\mu}_f = \bar{\mu}$
Partial Coverage $\hat{\mu}_f = \tilde{\mu}_f$
0
Λ
$\tilde{\lambda}(\tilde{r}_f)$
$\tilde{\lambda}(\mu^o)$
$\tilde{\lambda}(\tilde{r}_b)$
BP
Full Coverage
I: $\hat{\mu}_b = \mu^o$
II: $\hat{\mu}_b = \bar{\mu}$
Partial Coverage
$\hat{\mu}_b = \tilde{\mu}_b$
FFS
Full Coverage $\hat{\mu}_f = \bar{\mu}$
Partial Coverage $\hat{\mu}_f = \tilde{\mu}_f$
0
Λ
$\tilde{\lambda}(\mu^o)$
$\tilde{\lambda}(\tilde{r}_f)$
$\tilde{\lambda}(\tilde{r}_b)$
BP
Full Coverage
I: $\hat{\mu}_b = \mu^o$
II: $\hat{\mu}_b = \bar{\mu}$
Partial Coverage $\hat{\mu}_b = \tilde{\mu}_b$

(a) $t < \bar{t}$ or/and $B < \bar{B}$; $\bar{\mu} = \max\{\mu : \tilde{\lambda}(\mu) = \Lambda\}$

(b) $t \geq \bar{t}$ and $B \geq \bar{B}$; $\bar{\mu} = \max\{\mu : \tilde{\lambda}(\mu) = \Lambda\}$

Figure 3: The Comparison of the Equilibrium Outcomes: FFS vs BP

As illustrated in Figure 3, when $\Lambda \geq \max\{\tilde{\lambda}(\tilde{r}_f), \tilde{\lambda}(\mu^o)\}$, $\hat{\mu}_b(\hat{r}_b) = \bar{\mu}$ and $\hat{\mu}_f(\hat{r}_f) = \tilde{\mu}_f(\tilde{r}_f)$. In this case, the first statement of Proposition B7 implies that the results for the case under which the potential patients are partially covered uner both the FFS and BP schemes still hold. That is, the BP scheme dominates the FFS scheme in terms of the service quality and the patient welfare, while the FFS scheme is better in terms of the congestion level. However, when $\tilde{\lambda}(\tilde{r}_f) < \tilde{\lambda}(\mu^o)$ and $\Lambda \in [\tilde{\lambda}(\tilde{r}_f), \tilde{\lambda}(\mu^o)]$, $\hat{\mu}_b(\hat{r}_b) = \mu^o$ (see Figure 3(a)). Then the second statement of Proposition B7 implies that if the potential arrival rate Λ is relatively large (i.e., $\Lambda \in [\bar{\Lambda}, \tilde{\lambda}(\mu^o)]$), the results for the case under which the potential patients are partially covered uner both the FFS and BP schemes hold. While if the potential arrival rate Λ is relatively small (i.e., $\Lambda \in [\tilde{\lambda}(\tilde{r}_f), \bar{\Lambda}]$), the results for the case under which the potential patients are fully covered uner both the FFS and BP schemes hold.

Proof of Propositions in Appendix B

Proof of Proposition B3. When $\tilde{\lambda}(\tilde{r}_f) < \Lambda$, the results given in Proposition 2 are feasible. Therefore, $\hat{r}_f = \tilde{r}_f$ and $\hat{\mu}_f(\hat{r}_f) = \tilde{\mu}_f(\tilde{r}_f)$. When $\tilde{\lambda}(\tilde{r}_f) \geq \Lambda$, the results under the partial coverage scenatio are infeasible. The health care system will end up with the full coverage in equilibrium. Recall that the patient welfare under the full coverage is independent of r_f; see

(32). Therefore, the patient welfare remains the same for any given r_f under the full coverage. However, the funder will not be more generous than it has to be. Therefore, it will choose the smallest feabile reimbursement rate. We next show that $\hat{r}_f$, where $\tilde{\mu}_f(\hat{r}_f) = \bar{\mu}$, is the smallest one in the feasible set of the funder's problem under the full coverage. According to Proposition B1, the HCP's optimal service rate $\bar{\mu}$ under the full coverage satisfies $\tilde{\lambda}(\bar{\mu}) = \Lambda$. Since $\tilde{\lambda}(\hat{r}_f) = \tilde{\lambda}(\bar{\mu}) = \Lambda$, and $\tilde{\lambda}(r_f)$ is increasing in r_f, $\hat{r}_f$ is the smallest one that satisfies the full coverage requirement. Furthermore, because $\tilde{\lambda}(\hat{r}_f) = \Lambda \leq \tilde{\lambda}(\tilde{r}_f)$, $\tilde{r}_f \geq \hat{r}_f$. Thus, $\hat{r}_f\tilde{\lambda}(\hat{r}_f) \leq \tilde{r}_f\tilde{\lambda}(\tilde{r}_f) = B$, which implies that $\hat{r}_f$ also satisfies the budget constraint (33). Therefore, $\hat{r}_f$ is the feasible solution under the full coverage scenario. As $\hat{r}_f$ is the smallest one that satisfies the full coverage requirement, it is also the smallest one in the feasible set of the funder's problem under the full coverage. □

Proof of Proposition B6. When $\tilde{\lambda}(\tilde{r}_b) < \Lambda$, the results given in Proposition 2 are feasible. Therefore, $\hat{r}_b = \tilde{r}_b$ and $\hat{\mu}_b(\hat{r}_b) = \tilde{\mu}_b(\tilde{r}_b)$. When $\tilde{\lambda}(\tilde{r}_b) \geq \Lambda$, the results under the partial coverage are infeasible. The health care system will end up with the full coverage in equilibrium. As the patient welfare under the full coverage is independent of r_b (see (38)), the patient welfare remains the same for any given feasible r_b under the full coverage. However, the funder will not be more generous than it has to be. Therefore, under the full coverage it will choose the smallest feasible reimbursement rate. We next show that when $\Lambda \leq \tilde{\lambda}(\tilde{r}_b)$, $\hat{r}_b$, where $\tilde{\mu}_b(\hat{r}_b) = \bar{\mu}$, is the smallest feasible reimbursement rate. Since $\tilde{\lambda}(\hat{r}_b) = \tilde{\lambda}(\bar{\mu}) = \Lambda$ and $\tilde{\lambda}(r_b)$ is increasing in r_b, $\hat{r}_b$ is the smallest one that satisfies the full coverage requirement. Furthermore, as $\tilde{\lambda}(\hat{r}_b) = \Lambda \leq \tilde{\lambda}(\tilde{r}_b)$, $\hat{r}_b \leq \tilde{r}_b$. Thus, $\hat{r}_b\tilde{\lambda}(\hat{r}_b) \leq \tilde{r}_b\tilde{\lambda}(\tilde{r}_b) = B$, which implies that $\hat{r}_b$ also satisfies the budget constraint (39). As $\hat{r}_b$ is the smallest one that satisfies the full coverage requirement, it is also the smallest one in the feasible set of the funder's problem under the full coverage. □

Proof of Proposition B7. When $\tilde{\lambda}(\tilde{r}_f) < \Lambda \leq \tilde{\lambda}(\tilde{r}_b)$, the potential patients are partially covered under the FFS scheme and are fully covered under the BP scheme. We first consider the case when $\tilde{\lambda}(\tilde{r}_b) \geq \Lambda > \tilde{\lambda}(\tilde{r}_f) \geq \tilde{\lambda}(\mu^o)$. From Propositions B3 and B6, we know that in equilibrium, $\hat{\mu}_f(\hat{r}_f) = \tilde{\mu}_f(\tilde{r}_f)$, $\hat{\mu}_b(\hat{r}_b) = \bar{\mu}$, and $\hat{\lambda}(\hat{r}_f) = \tilde{\lambda}(\tilde{r}_f) < \Lambda = \hat{\lambda}(\hat{r}_b)$. Because $\tilde{\lambda}(\mu)$ is unimodal in μ, $\bar{\mu}$ is the larger root of $\tilde{\lambda}(\mu) = \Lambda$, $\frac{d\tilde{\lambda}(\tilde{\mu}_f(\tilde{r}_f))}{d\mu} < 0$ (shown in the proof of Proposition 3) and $\tilde{\lambda}(\tilde{r}_f) \leq \Lambda$, $\bar{\mu} < \tilde{\mu}_f(\tilde{r}_f)$. Hence, $\delta(\hat{r}_b) < \delta(\hat{r}_f)$. Since both $W(\mu)$ and $T(\mu)$ are decreasing in μ (see (26)), $W(\hat{r}_f) = W(\tilde{r}_f) < W(\bar{\mu}) = W(\hat{r}_b)$ and $T(\hat{r}_f) = T(\tilde{r}_f) < T(\bar{\mu}) = T(\hat{r}_b)$. As $n(\mu)$ increases in μ, $n(\hat{r}_b) < n(\hat{r}_f)$. When patients are fully covered, $U(\Lambda, \hat{\mu}_b(\hat{\mu}_b)) > 0$. Based on (14) and (38), we have $S(\hat{r}_f) = -\beta(\Lambda - \tilde{\mu}_f(\tilde{r}_f)) < 0 < \Lambda \cdot U(\Lambda, \hat{\mu}_b(\hat{\mu}_b)) = S(\hat{r}_b)$.

We next consider the scenario $\tilde{\lambda}(\tilde{r}_f) < \tilde{\lambda}(\mu^o)$. First, when $\tilde{\lambda}(\tilde{r}_b) \geq \Lambda > \tilde{\lambda}(\mu^o)$, from Propositions B3 and B6, $\hat{\mu}_f(\hat{r}_f) = \tilde{\mu}_f(\tilde{r}_f)$, $\hat{\mu}_b(\hat{r}_b) = \bar{\mu}$, and $\hat{\lambda}(\hat{r}_f) = \tilde{\lambda}(\tilde{r}_f) < \Lambda = \hat{\lambda}(\hat{r}_b)$.

Similar to the above analysis, we can show that $\delta(\hat{r}_b) < \delta(\hat{r}_f)$, $W(\hat{r}_f) < W(\hat{r}_b)$, $T(\hat{r}_f) < T(\hat{r}_b)$ and $S(\hat{r}_f) < S(\hat{r}_b)$.

When $\Lambda \in [\tilde{\lambda}(\tilde{r}_f), \tilde{\lambda}(\mu^o)]$, utilizing Propositions B3 and B6 again, we have $\hat{\mu}_f(\hat{r}_f) = \tilde{\mu}_f(\tilde{r}_f)$, $\hat{\mu}_b(\hat{r}_b) = \mu^o$, $\hat{\lambda}(\hat{r}_f) = \tilde{\lambda}(\tilde{r}_f) < \Lambda = \hat{\lambda}(\hat{r}_b)$. Because $\bar{\mu}$ is the larger solution of $\tilde{\lambda}(\mu) = \Lambda$, $\tilde{\lambda}(\mu)$ is unimodal in μ and $\Lambda \leq \tilde{\lambda}(\mu^o)$, $\mu^o < \bar{\mu}$. We have shown above that $\bar{\mu} < \tilde{\mu}_f(\tilde{r}_f)$. Thus, $\hat{\mu}_b(\hat{r}_b) = \mu^o < \tilde{\mu}_f(\tilde{r}_f) = \hat{\mu}_f(\hat{r}_f)$ and $\delta(\hat{r}_b) < \delta(\hat{r}_f)$. Based on (14) and (38), we have $S(\hat{r}_f) = -\beta(\Lambda - \tilde{\mu}_f(\tilde{r}_f)) < 0 < \Lambda \cdot U(\Lambda, \hat{\mu}_b(\hat{\mu}_b)) = S(\hat{r}_b)$.

Finally, we show that there exists a $\bar{\Lambda}$ such that $T(\hat{r}_b) < T(\hat{r}_f)$ iff $\Lambda < \bar{\Lambda}$. First, because potential patients are fully covered under the BP scheme but are partially covered under the FFS scheme, from (6) and (26), we get

$$\begin{aligned} T(\hat{r}_b) &= T(\Lambda, \hat{\mu}_b(\hat{r}_b)) = T(\Lambda, \mu^o) = \frac{1}{o(\mu^o) - \Lambda}, \\ T(\hat{r}_f) &= T(\tilde{\lambda}(\tilde{r}_f), \tilde{\mu}_f(\tilde{r}_f)) = \frac{R}{\theta} - \frac{t}{\theta(1 - \delta(\tilde{r}_f))}, \end{aligned}$$

where $T(\hat{r}_b)$ is increasing in Λ while $T(\hat{r}_f)$ is independent of Λ. Therefore, if there exist Λ_1 and Λ_2 such that $T(\hat{r}_b) = T(\Lambda_1, \hat{\mu}_b(\hat{r}_b)) > T(\hat{r}_f)$ and $T(\hat{r}_b) = T(\Lambda_2, \hat{\mu}_b(\hat{r}_b)) < T(\hat{r}_f)$, then there must exist a $\bar{\Lambda}$ such that $T(\hat{r}_b) < T(\hat{r}_f)$ iff $\Lambda < \bar{\Lambda}$. Denote $\Lambda_1 = \tilde{\lambda}(\mu^o)$ and $\Lambda_2 = \tilde{\lambda}(\tilde{r}_f)$. Then, when $\Lambda = \Lambda_1$, we can show that

$$T(\hat{r}_b) = T(\tilde{\lambda}(\mu^o), \mu^o) = \frac{R}{\theta} - \frac{t}{\theta(1 - \delta(\mu^o))} > \frac{R}{\theta} - \frac{t}{\theta(1 - \delta(\tilde{r}_f))} = T(\hat{r}_f),$$

where the inequality is due to that $\delta(\mu)$ is increasing in μ and $\mu^o < \tilde{\mu}_f(\tilde{r}_f)$. When $\Lambda = \Lambda_2$, based on (6) and (8), we can show that

$$T(\hat{r}_f) = \frac{1}{o(\tilde{r}_f) - \tilde{\lambda}(\tilde{r}_f)} > \frac{1}{o(\mu^o) - \tilde{\lambda}(\tilde{r}_f)} = T(\tilde{\lambda}(\tilde{r}_f), \hat{\mu}_b(\hat{r}_b)) = T(\hat{r}_b),$$

where the inequality is due to that $o(\mu)$ achieves its maximum at μ^o. We thus complete the proof. □