UNIVERSITY OF
CAMBRIDGE
Judge Business School

Working Paper Series

3/2011

Performance-Based Contracts for Outpatient Medical Services

Houyuan Jiang, Zhan Pang and Sergei Savin

Cambridge Judge Business School Working Papers

Cambridge Judge Business School author contact details are as follows:

Houyuan Jiang
Cambridge Judge Business School
University of Cambridge
h.jiang@jbs.cam.ac.uk

Please address enquiries about the series to:

Research Manager
Cambridge Judge Business School
Trumpington Street
Cambridge CB2 1AG
UK

Tel: 01223 760546
Fax: 01223 339701
Email: research-support@jbs.cam.ac.uk

Performance-Based Contracts for Outpatient Medical Services

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

Zhan Pang
Lancaster University Management School,
Lancaster, LA1 4YW, United Kingdom, z.pang@lancaster.ac.uk

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

In recent years, the performance-based approach to contracting for medical services has been gaining popularity across different healthcare delivery systems, both in the US (under the name of "Pay-for-Performance", or P4P), and abroad ("Payment-by-Results", or PbR, in the UK). One common element of performance-based compensation is the inclusion of patient service access metrics, in addition to the quality of clinical outcomes, in the process of performance evaluation for a provider of healthcare services. For example, the implementation of the "Payment-by-Results" approach includes appointment scheduling targets designed to shorten patient waiting time, and adherence to these targets is monitored through a dedicated online outpatient appointment system, "Choose-and-Book".

The goal of our research is to build a unified performance-based contracting (PBC) framework that incorporates patient access-to-care requirements and that explicitly accounts for the complex outpatient care dynamics facilitated by the use of an online appointment scheduling system. In our model, a service provider needs to allocate his service capacity among three patient groups: urgent patients whose service cannot be postponed, and two groups of non-urgent patients, dedicated patients who insist on getting served by their first-choice provider and flexible patients who will choose another provider if the online appointment system shows no available appointments with their first-choice provider. The principal wants to minimize her cost (payments made to the provider offset by the waiting-time penalty) of achieving the expected waiting-time target. We model the appointment dynamics in the presence of a mixed-patient population as that of an $M/D/1$ queue and analyze several contracting approaches under adverse selection (asymmetric information) and moral hazard (private actions) settings. We study the first-best and the second-best solutions, as well as their specific contracting implementation schemes. Our results show that simple and popular schemes used in practice cannot implement the first-best solution and that the linear PBC cannot implement the second-best solution. In order to overcome these limitations, we propose a threshold-penalty PBC approach and show that it coordinates the system for an arbitrary patient mix and that it achieves the second-best performance for the setting where all patients are dedicated.

Key words: Healthcare; performance-based contracting, principal-agent theory; queueing theory.

1. Introduction

As the US healthcare system is preparing to face a set of fundamental changes, the overarching task of controlling the cost of providing medical care while maintaining a high quality and a satisfactory level of access to care occupies one of the central places in current political debate. The evidence that continuing increases in healthcare spending in many instances do not translate into desired improvements in quality of care or into better patient outcomes (Institute of Medicine Report (2001), McGlynn *et al.* (2003), Fisher *et al.* (2004), Leape and Berwick (2005)) suggests that reform of the overall healthcare system should include significant changes in the current mechanisms of compensating healthcare providers for the services they deliver. In the domain of publicly financed healthcare programs, Medicare, which leads both in terms of the number of patients covered and financial spending, provides an important example of the historical evolution of these compensation practices. Initially, Medicare employed a retrospective payment approach under which service providers (hospitals and physicians) received compensation based on the minimum of usual and customary charges or the actual costs incurred while delivering a particular service. The exact payment amount was known only at the end of the calendar period, after the customary charges were established and the incurred costs verified. The incentives to perform more services and to charge more for them were built into this payment system and led to a gradual escalation of costs and charges, which, in turn, resulted in the introduction in 1983 of the current, prospective payment mechanism. Unlike the retrospective approach, the prospective system uses a schedule of predetermined fees that are calculated for each particular kind of medical service by the government body, Centers for Medicare and Medicaid Services (CMS) [1]. The intended purpose of the fee-for-service (FFS) system was to incentivize providers to improve the efficiency of their service delivery processes and to contain their per-unit-of-service costs at the levels specified by fee-for-service schedules. Although an improvement over the retrospective system in terms of limiting the growth of healthcare expenditures, the FFS approach still suffers from major weaknesses. Firstly, it still encourages providers to increase the volume of provided services and shifts providers' attention towards new, complex and more expensive treatments and away from less costly alternatives. In addition, FFS payments are not tied to the quality of provided services as measured by patient experiences and clinical outcomes, and thus do not provide any incentives for preventive activities or for coordination of patient care, that may reduce the need for future costly interventions. These and other limitations of the fee-for-service approach are summarized in the seminal Institute of Medicine

[1] At the end of 2009, 77% of the total of 45 million Medicare patients were enrolled in the federally administered fee-for-service program (Kaiser Family Foundation Report (2009)).

2006 Report, “Rewarding Provider Performance: Aligning Incentives in Medicare”, which calls for the introduction of an alternative, “Pay-for-Performance” (P4P) provider compensation scheme. Under the P4P compensation approach, not only the quantity but also the quality of provided services influence the compensation amounts. In all cases of P4P adoption, the reported quality metrics include prophylaxis measures, such as cancer screenings and coordinated diabetes care, as well as clinical outcomes, such as hospital readmission rates and preventable hospitalizations (Mullen *et al.* (2010)). In a number of cases, clinical performance measures are augmented by “patient experience” metrics that include prompt access to care (Integrated Healthcare Association (2010)).

While P4P framework is only now emerging from its pilot status in the US, it is a well accepted paradigm in a number of European countries as well as in Australia. In the United Kingdom, in particular, it is already used at the national level by the National Health Service (NHS), which coordinates both the financing and the delivery of healthcare services. Since 2002, the NHS uses a system of hospital financing called “Payment-by-Results”, or PbR (since 2004 this system has also been applied to primary care physicians) [2]. Similar to the fee-for-service approach adopted by Medicare, PbR ensures that a service provider (e.g., a hospital) receives a fixed payment from a service purchaser (a government agency) for each delivered treatment, with the payment amount determined by the treatment’s Health Resource Group (HRG) code (the UK analog of the US diagnosis-related group, or DRG, code). Under the PbR system, primary care trusts (PCTs, the commissioning agencies of the NHS) are free to purchase healthcare services from any qualified local provider, in either the public or private sector. Unlike fee-for-service, PbR includes various service quality measures, including those related to patient access to care. In particular, the NHS currently uses a series of patient waiting-time targets including the 18-week period as a maximum waiting time for any outpatient to receive elective specialist care [3] (most specialist care in the UK is done in state-managed hospitals). Overall, the PbR mechanism ties a substantial portion of physician/hospital compensation (as much as 18% (Roland (2004)) to conformance with service quality standards [4]. A representative example of how patient waiting times influence provider compensation is provided by the 2008 standard NHS contract for acute services (NHS Contract (2008)), which stipulates penalties of up to 5% of the revenue from elective services for violating the 18-week waiting target. Recently, in order to facilitate better patient access to care and to

[2] http://www.dh.gov.uk/en/Managingyourorganisation/Financeandplanning/NHSFinancialReforms/index.htm

[3] http://www.18weeks.nhs.uk

[4] The corresponding figure for the US P4P initiatives is about 5% (Rosenthal *et al.* (2004)).

streamline the management of outpatient appointments, a nation-wide electronic appointment booking system, Choose-and-Book (CaB), was set up.

While these innovations are actively changing the way healthcare delivery systems operate, the nature of interactions between different contractual obligations imposed on service providers remains poorly understood. The goal of our research is to build a unified performance-based contracting (PBC) framework that incorporates patient access-to-care requirements and that explicitly accounts for the complex outpatient care dynamics facilitated by the use of an online appointment scheduling system. Our model of outpatient care is based on the UK setting, where a hospital, based on private information about its operational costs, makes two types of capacity allocation decisions: how many appointment slots to make available through the online appointment scheduling system (and, consequently, how many to reserve for same-day urgent cases), and how many days in advance to release such capacity into the online system (CaB). Using these two decision levers, the hospital allocates its service capacity between same-day patients as well as two distinct types of patients with delayed service requests, "dedicated" and "flexible". Dedicated patients insist on having their service provided by a particular hospital, irrespective of whether the CaB system shows any appointments available in that hospital - and they have the recourse to enforce an appointment within the 18-week horizon through the use of a phone-based override system. Flexible patients, on the other hand, will select another service provider and forgo the additional inconvenience associated with using the override if the CaB system displays no available appointments within the horizon selected by their first-choice provider. We assume that the hospital receives a known revenue from the government agency (similar to an FFS payment) for each patient receiving care. In addition, the hospital incurs penalties if its operational planning turns out to be inadequate. First, the overtime penalty is incurred in cases when the total daily demand for outpatient services exceeds the hospital's nominal service capacity (the value of the overtime cost is assumed to be the hospital's private information). Also, every time a patient switches to another hospital due to lack of appointment capacity as declared through the CaB system, a "work transfer" penalty is incurred. Finally, the government agency charges the hospital an "access-to-care" penalty proportional to the length of the appointment waiting list it has. The revenue amount and the access-to-care penalty value form the core of the hospital's PBC put forward by the government agency. In our analysis, we consider an asymmetric information setting in which a hospital has perfect knowledge about the value of its overtime costs while the government knows only the distribution of its potential values. In this setting, the government agency makes its decision regarding the parameters of the PBC in anticipation of the provider's rational choices with respect to the proportion of the total daily capacity allocated to advanced appointments and the threshold on the queue of advance appointments.

Such a contract can be modeled using the principal-and-agent framework in which the purchaser of services acts as a principal and the service provider as an agent. The agent aims to maximize his profits, which consist of the payment for provided services net of the penalty for making patients wait and the overtime and the work-transfer cost. The principal wants to minimize her cost (payments made to the provider offset by the waiting-time penalty) of achieving the expected waiting-time target. In addition, the principal's problem includes the standard individual rationality constraint inducing the agent to accept the contract, as well as the incentive compatibility constraint forcing the agent to reveal the true value of her overtime cost. Using this principal-agent framework, we analyze both the FFS and PBC approaches under adverse selection (asymmetric information) and moral hazard (private actions) settings. For both settings, we study the first-best and the second-best solutions, as well as the performance of a simple contract that applies the same contract parameters to all agents, irrespective of their overtime cost values. In particular, our analysis addresses the following questions:

- What is the optimal structure for each type of contract under different information settings?
- When does the PBC approach result in better outcomes for the principal; and,
- What is the impact of the waiting-time target on the agent's decisions and the principal's optimal contract design?

In our analysis, we gain important insights by comparing the FFS and PBC mechanisms in different settings: with complete information, with asymmetric information, and with private agent actions. In particular, we show that when the agent's capacity allocation decisions are observable and contractible, the FFS and PBC approaches produce the same outcome, irrespective of whether the information setting is symmetric or asymmetric; see Proposition 2 and Proposition 4, respectively. However, if the agent's decisions are not observable and contractible, PBC outperforms FFS. This suggests that PBC should replace FFS in settings similar to the one observed in the UK NHS system, where the government does not routinely collect operational cost information and where hospitals possess a lot of power for making their own capacity management decisions.

The rest of this paper is organized as follows. Section 2 reviews the related research. Section 3 describes our model in detail. Sections 4 and 5 analyze the setting in which all of the capacity allocation decisions are made by the purchaser of services (principal) under either complete or incomplete information. In Section 6, we consider threshold penalty performance-based contracts, which can achieve the first-best outcome for any diverting rate, and for the special case of having dedicated patients only, which can achieve the second-best outcome. We conclude the paper in Section 7.

2. Literature Review

Goddard *et al.* (2000) and Farrar *et al.* (2007) describe conceptual frameworks for designing fee-for-service contracts from an economic perspective and outline potential risk factors associated with the FFS approach, in particular, decreased quality of delivered services and reduced access to care. Farrar *et al.* (2007) present an empirical study of the consequences of introducing FFS schemes in the UK and provide evidence of reductions in the cost of care. At the same time, there appears to be less evidence of increased volume of delivered services, and no evidence of a detrimental impact on the quality of care. De Fraja (2000) underscores the information asymmetry between a purchaser of services (government agency) and a service provider (hospital) inherent in healthcare settings and presents a stylized model of FFS contracting based on the principal-and-agent framework. In particular, it is shown that in settings where providers have private information about their costs, lower-cost providers may receive, under the optimal contract, a higher compensation per unit of delivered services.

In the UK, the performance of physicians and hospitals on a number of clinical and patient access-to-care metrics is measured vigorously, and performance violations lead to financial penalties and can threaten the careers of those who manage the delivery of care. It is only natural that the PbR, originally conceived as an activity-based, FFS mechanism, has gradually evolved to include provider performance metrics. Contract theory literature streams in economics and operations management (see Bolton and Dewatripont (2005) and Cachon (2003) for comprehensive reviews) include a large number of papers that focus on designing incentives to induce desired performance. Below we highlight several studies on service supply chain contracting which are closely related to our work.

In the call-center context, Ren and Zhou (2008) and Hasija *et al.* (2008) study coordination mechanisms in the setting where a client company outsources call-center operations to a vendor. Ren and Zhou (2008) model call-center operations using a fluid approximation to a $G/G/s$ queue with customer abandonment and present a principal-agent model in which the agent (vendor) controls the staffing level (service capacity) and the level of effort focused on achieving the desired service quality (defined as a fraction of customer calls successfully "resolved"), and the principal focuses on coordinating these decisions using a family of performance-based and cost-sharing contracts. It is shown that neither the fee-for-service, or "piecemeal", contract nor its performance-based extension under which only "resolved" calls are compensated, manage to achieve the first-best solution. At the same time, system coordination is achieved, irrespective of whether the agent's quality effort level is observable to the principal and contractible or not, if the performance-based contract is augmented by an appropriate cost-sharing mechanism. Hasija *et al.* (2008), on the other hand, model a call center as an $M/M/N$ queue with customer abandonment and use a diffusion approximation derived in Garnett *et al.* (2002) to derive service performance measures, such as the expected waiting time, the probability of the waiting time exceeding a given threshold, and the equilibrium probability of abandonment. In a principal-agent setting similar to the one described in Ren and Zhou (2008), the principal uses a family of contracts which include FFS (pay-per-time (PPT) and pay-per-call (PPC)) as well as PBC (service-level agreement (SLA) constraint and associated penalty, and waiting-time penalty) elements, and the agent maximizes her expected profit by making capacity sizing and productivity (service rate) decisions. An important feature of the proposed model is the information asymmetry between the principal and the agent with respect to the agent's productivity value. The authors establish, in particular, that a combination of a PPC and a PPT-based contract which incorporates linear penalties for customer waiting time coordinates the system and allows the principal to maximize her profit without paying any information rent. While information asymmetry generally produces an information rent for the agents with the information advantage, in the model presented in Hasija *et al.* (2008), for a given capacity decision, the asymmetric information about productivity levels impacts only the principals revenue and the transfer payments but not the agents' costs. Under a coordinated solution, both the capacity level and the transfer payment for the high-productivity agent is lower than that for the low-productivity agent, while the marginal revenue rate per customer for the high-productivity agent is higher than that for the low-productivity agent. This explains why it is possible for the principal to avoid paying information rent by offering a PPC-based contract to the high-productivity agent and a PPT-based contract to the low-productivity one.

Similar to Ren and Zhou (2008) and Hasija *et al.* (2008), we examine the role of the activity-based and the performance-based incentives on the structure of service contracts. Despite the similarity of a research agenda, our modeling approach differs from the one adopted in Ren and Zhou (2008) and Hasija *et al.* (2008) in several essential ways, reflecting the reality of a typical outpatient setting. Firstly, our model explicitly treats outpatient appointment and service dynamics as that of an $M/D/1$ queue without using first-moment or diffusion approximations; this feature, however, is slightly moderated by the fact that we ignore customer abandonments. Secondly, the information structure of our model is different from that of Hasija *et al.* (2008) (Ren and Zhou (2008) do not analyze information asymmetry). In particular, Hasija *et al.* (2008) consider information asymmetry in agents' service rates. In the outpatient care setting we model, provider productivity is visible to the purchaser of healthcare services, and the most important aspect of the information asymmetry concerns the provider's overtime costs. As a result, in their model the principal can design a contract to eliminate the entire information rent, while in our model the information rent is unavoidable. Thirdly, and most importantly, in our model the agent's decisions shift from capacity sizing and effort/productivity level management in the face of a homogenous customer base to the rather different task of allocating fixed service capacity among three different patient groups.

In the context of the after-sales service supply chains for complex, multi-component products, Kim *et al.* (2007) introduce a multi-task principal-agent model to analyze contracts observed in practice. Their model describes a setting in which a principal (the end-user of the product) is faced with the task of coordinating the inventory stocking and the cost-cutting effort level decisions of multiple agents, each of which is responsible for supplying one of the product's essential components. The expected product downtime is used as a performance metric, and it is shown that while the performance-based contract achieves the first-best solution in the setting where all channel members are risk-neutral, an additional cost-sharing mechanism is required to produce the second-best outcome in the setting where channel members are risk-averse. In addition to two features described above (use of the $M/D/1$ queueing approach to model the outpatient appointment and service dynamics and the capacity allocation nature of the agent's problem), another important difference between our work and Kim *et al.* (2007) is the type of modeling assumption that generates the inefficiency of basic performance-based contracting approaches. In our model, both parties are risk-neutral, but there exists an information asymmetry between them, while in Kim *et al.* (2007) the same information is available to the principal and the agents, and both parties are risk-averse.

The number of applications of contract theory to healthcare services, while somewhat limited as compared to retail and other service supply chains, has been growing in recent years, in part due to the increased popularity of the performance-based contracts. Lu and Donaldson (2000) present a review of the economics literature dealing with performance-based contracting, and underscore that the inherent informational advantage that healthcare providers have over patients as well as agencies purchasing services is one of the major sources of potential market failure in the healthcare domain [5]. Under a dynamic principal-agent framework, Fuloria and Zenios (2001) study an

[5] An interesting exception to this general statement is analyzed in Su and Zenios (2006), where, in the kidney transplantation context, patients may have an informational advantage over care providers.

outcomes-adjusted payment system where the purchaser determines the contract terms contingent on the observed outcomes (patient deaths and medical complications) to induce the provider to choose the optimal treatment intensities. Our analysis differs from the one presented in Fuloria and Zenios (2001) in several ways. First, we focus on the operational performance measure (patient waiting time) rather than on the clinical outcomes. Second, while Fuloria and Zenios (2001) consider only the moral hazard setting, we analyze an information asymmetry setting which leads to both moral hazard and adverse selection. Finally, Fuloria and Zenios (2001) focuses on the linear contract structure, while we also study non-linear, threshold performance-based contracts. Lee and Zenios (2007) study evidence-based incentive systems within a multi-task principal-agent model in the context of dialysis treatment for patients with end-stage renal disease. In particular, they develop an empirical methodology to identify appropriate intermediate performance measures to be included in the overall agent's performance set in addition to outcome-based metrics, such as patient hospitalization frequency. In the outpatient service context analyzed in our work, most of the performance measures used in practice (such as patients' waiting time) fall within the category of intermediate performance measures. So and Tang (2000) consider a Medicare contract for the reimbursement of drug prescriptions in an outpatient environment with a clinical outcome-based performance metric and derive the optimal drug application policy that maximizes the outpatient clinic's expected profit. This paper, however, does not analyze the optimal contract structure nor does it impose, due the context of the analyzed problem, a limit on outpatient clinic service capacity. A separate research stream within the healthcare contracting literature focuses on the issues of excess demand and waiting for service (for a comprehensive review see Siciliani (2007)). While several existing papers model the information asymmetry between the purchasers of services and their providers, none of them analyze the underlying service capacity management issues and their impact on patient waiting times.

In our model, we use a principal-agent set-up in which the agent solves the problem of allocating its service capacity among the same-day patients and the patients who use the online appointment system. The extant literature contains numerous papers that deal with various instances of service capacity allocation in healthcare settings (for example, see Gupta and Denton (2008) for a comprehensive review of recent advances in the appointment scheduling literature). However, to the best of our knowledge, our work is the first to incorporate appointment capacity allocation within the contracting principal-agent framework.

3. Contracting for Outpatient Medical Services: The Model

We consider a healthcare service contract problem in which a purchaser of services (a government agency, such as a primary care trust in the UK context) offers a contract to a provider (a hospital) to deliver outpatient services. The provider manages outpatient appointments via an online outpatient appointment booking system (such as Choose-and-Book). Demand for outpatient services is random and is comprised of two distinct streams: advance appointments which can be served either on the current day or on a future date, and same-day appointments which must be served on the day they arrive. The provider has a limited nominal daily service capacity, but is obligated to serve all same-day appointments and all accepted advance appointments due on each day; when the total number of patients requiring service on a particular day exceeds the nominal daily service capacity, the provider incurs overtime costs to cover the extra demand.

A waiting list (queue) for advance appointments arises as a result of uncertain demand and limited service capacity. The provider manages its limited service capacity under an incentive structure that includes a fixed revenue for serving each patient (a fee-for-service component) and penalties for delaying or refusing patient service (the performance-based component). The purchaser of services needs to minimize the service cost while meeting an appointment waiting time target. The interaction between the purchaser and the provider can be recast as a principal-agent model, where the purchaser acts as a principal and the provider as an agent.

3.1. Capacity Allocation Policy, Appointment Backlog Dynamics and Cost Structure

We assume that the provider has a nominal capacity of C equal-length outpatient time slots per day. Advance appointment requests from the online appointment booking system (CaB) arrive according to a Poisson process with an average daily demand rate of λ (arrivals on different days are independent). Advance appointments are divided into two classes: dedicated and flexible. A dedicated patient makes an appointment either through the CaB if she finds an available time slot or through the override phone-based system (in the UK, the national Telephone Advice Line, TAL) if no time slot is available from her chosen provider through the CaB. In other words, a dedicated patient insists on being serviced by her first-choice provider for reasons of geographical proximity, provider's reputation, etc., even if this may result in a longer wait and extra administrative costs in getting an appointment through the override route. A flexible patient, on the other hand, is unwilling to incur the extra cost associated with the override option and makes an appointment with another provider if the CaB shows no appointment available with his first-choice provider. We assume that a patient who finds that no appointment slots are available through the CaB with the first-choice provider turns out to be a dedicated patient with probability θ, a parameter which describes the perceived level of provider reputation/popularity. In particular, $\theta = 1$ describes a unique facility with a strong reputation for a particular kind of specialist services, while $\theta = 0$ would characterize an undistinguished facility with easy-to-find substitutes. In reality, patients may make choices not only in terms of preferred providers, but also in terms of the day and time slots on which they would like to be seen. For tractability, we assume that advance-booking patients always choose the earliest appointment time slot available through the CaB. We also ignore the phenomenon of no-shows and assume that all patients punctually show up for their appointments. The same-day demand for outpatient services, D_0, is assumed to be a discrete random variable with cdf $F_{D_0}(\cdot)$, statistically independent from the demand for advance appointments. We also assume that, each day, same-day patients are served after advance appointments.

Hospital management is faced with the problem of allocating its service capacity among three patient groups: advance dedicated, advance flexible, and same-day patients. While every hospital in the UK is required to manage its advance appointments using the CaB system, the exact fraction of its service capacity to be released to the CaB is within a hospital's discretion. We consider the following (A, Z) capacity allocation policy: the hospital releases to CaB A out of C daily appointment slots starting from the present day until some time in the future so that the total number of released slots is equal to Z. This policy ensures that $C - A$ daily appointment slots are reserved for same-day patients, and that the flexible advance demand is blocked from entering the system when the appointment backlog exceeds Z/A days. Our interactions with hospital managers in several UK hospitals confirm the appeal of this type of capacity allocation approach due to the simplicity and ease of its implementation. We assume that, with very high probability, the length of appointment backlog exceeds A slots, or, in other words, that patients almost always wait for their appointments for more than one day. Such an assumption was satisfied in all outpatient appointment environments we observed, and allows us to model the evolution of the appointment backlog under (A, Z) policy as that of an $M/D/1$ queue, where D reflects the fixed duration of an appointment slot, and the single-server feature describes the patient service dynamics proceeding at the rate of A slots per day. Note that the patient appointment backlog grows both during and outside of office hours, since appointment requests can arrive to the CaB system at any point during the day. At the same time, appointment backlog reduction can happen only during the part of the day corresponding to A slots, and during the rest of the day no appointment patients are served. While such a dynamics is best described using the framework of queues with server vacations (Tian and Zhang (2006)), no closed-form expressions for queueing performance measures exist within this framework. Instead, we assume that the server works continuously, and that the entire demand for appointments arrives only during the time period corresponding to A slots, at the rate of λ/A per slot if the appointment backlog is smaller than Z, and of $\theta\lambda/A$ if the appointment backlog is equal to or larger than Z (when no slots are available through the CaB, only dedicated patients can get appointments). Note that this queueing system, which we denote as *modified* $M/D/1$ *queue*, reduces to a standard $M/D/1$ queue when $\theta = 1$, and to a finite-buffer $M/D/1/Z$ queue when $\theta = 0$. To ensure the stability of the appointment backlog system, we assume that the minimum offered load value, $\theta\lambda/A$, is strictly less than one, which implies that $\theta\lambda < A$. Let $X(A,Z)$ be the random variable denoting the number of appointments in the system under the capacity allocation policy (A,Z). Then, the expected daily number of diverted patients is $\lambda(1-\theta)Pr(X(A,Z) \geq Z)$, and the expected daily throughput for advance appointments is $\lambda(1-(1-\theta)Pr(X(A,Z) \geq Z))$. Note that since the stationary distribution of $X(A,Z)$ depends on A only through the value of the offered load, we can treat A as a continuous variable in our analysis.

Hospital operational cost structure includes three terms: fixed maintenance and labor costs, which we normalize to zero, the cost for diverting patients, and the overtime costs. The patient-diversion cost represents the effect of the loss of goodwill for refusing to serve flexible patients and forcing them to select another care provider. If b is the cost for diverting one patient, the expected daily diverting cost is given by

$$P(A,Z) = b\lambda(1-\theta)Pr(X(A,Z) \geq Z). \tag{1}$$

Information asymmetries between purchasers and providers and between providers and clients pervade healthcare service supply chains (Arrow (1963), Haas-Wilson (2001), Bloom *et al.* (2008)). It is often pointed out that providers occupy a unique position in healthcare service delivery systems since they have an informational advantage over both the purchasers of services and the patients. In healthcare settings, the cost structure of the service provider often remains private knowledge which is neither communicated to nor verified by the purchaser. In particular, in an outpatient care environment, the key component of this private cost structure is the overtime cost that reflects the provider's ability to stretch its daily service capacity to match unexpected surges in same-day patient demand. In our analysis, we assume that the value of this cost constitutes private information, and we consider a setting in which the hospital's overtime cost associated with serving one patient can take two values, o^t, $t \in \{H,L\}$, with $o^H > o^L$. The hospital is said to be of type H if $t = H$; otherwise, it is said to be of type L. Under the assumption that the length of the appointment backlog almost always exceeds A, the expected daily overtime cost for the hospital of type t is given by

$$O^t(A) = o^t E_{D_0}[(D_0 - C + A)^+]. \tag{2}$$

Under the information asymmetry setting we model, the purchaser of services only knows the distributional information regarding the value of the provider's overtime cost: for the purchaser, the hospital is of type H (L) with probability $0 < p < 1$ (respectively, $1 - p$).

3.2. Performance Metric: Patient Waiting Time

The set of measures used in practice for evaluating the performance of healthcare providers include clinical outcomes as well as other quality-of-service metrics. In our analysis, we concentrate on the expected waiting time for advance appointments (expressed in terms of the number of appointment slots), $W_q(A,Z)$, as a measure of patient access to care. Let $L_q(A,Z) = E[(X(A,Z) - 1)^+]$ be the expected length of the waiting list. Since the expected value of the effective daily demand for appointments is $\lambda(1 - (1 - \theta)Pr(X(A,Z) \geq Z))$, it follows from Little's law that

$$L_q(A,Z) = \lambda(1 - (1 - \theta)Pr(X(A,Z) \geq Z))\frac{W_q(A,Z)}{A}. \tag{3}$$

For general values of provider reputation factor θ and arbitrary capacity allocation policy (A,Z), there exist no closed-form expressions for $L_q(A,Z)$ or $W_q(A,Z)$. However, it is possible to derive monotonicity properties for these quantities, which are helpful in analyzing performance-based contracts.

Proposition 1. *For the modified $M/D/1$ queue,*

(a) $L_q(A,Z)$, $W_q(A,Z)$, and $W_q(A,Z)/A$ are monotone increasing in θ, λ, and Z, and monotone decreasing in A, and

(b) $Pr(X(A,Z) \geq Z)$ is monotone increasing in θ and λ, and monotone decreasing in A and Z.

Most of the results of Proposition 1 are intuitive. On the one hand, an increase in θ, or λ or Z indicates an increase in the effective demand for appointments, which implies an increase in average queue length and in average waiting time expressed in terms of the number of appointment slots or working days. On the other hand, an increase in A indicates an increase in service capacity, which implies a decrease in average queue length and in average waiting time expressed in terms of the number of appointment slots or working days. The monotonicity property of $Pr(X(A,Z) \geq Z)$ with respect to the value of the CaB booking limit Z is, however, less obvious. As Proposition 1 indicates, the probability that the queue length exceeds the chosen CaB limit decreases with the value of that limit.

In the special cases of $\theta = 1$ and $\theta = 0$, waiting-time performance measures can be expressed in closed form under any (A, Z) policy. In particular, for a hospital with an entirely dedicated patient population ($\theta = 1$), the offered load is $\rho = \lambda/A$, and the Pollaczek-Khintchine result implies that

$$L_q(A,Z) = \frac{\rho^2}{2(1-\rho)} = \frac{\lambda^2}{2A(A-\lambda)}, \tag{4}$$

and

$$W_q(A,Z) = \frac{\rho}{2(1-\rho)} = \frac{\lambda}{2(A-\lambda)}. \tag{5}$$

Note that since during a working day the number of advance appointment slots is A, the expected number of *days* a patient has to wait is

$$\frac{W_q(A,Z)}{A} = \frac{\lambda}{2A(A-\lambda)}. \tag{6}$$

On the other hand, if all patients are flexible ($\theta = 0$), the appointment dynamics under the (A, Z) policy corresponds to that of a finite-buffer $M/D/1/Z$ queue. Closed-form expressions for the stationary distribution for such system are presented in Brun and Garcia (2000).

The principal operates under the constraint on the maximum value of the expected waiting time:

$$\frac{W_q(A,Z)}{A} \le M, \tag{7}$$

where M is waiting-time target measured in days. Note that for $\theta = 1$ or $Z = \infty$, $W_q(A,Z)/A = \frac{\lambda}{2A(A-\lambda)}$. From the result of Proposition 1 it follows that the service-level constraint (7) implies a lower bound for the value of A,

$$A^* = \frac{\lambda}{2} + \sqrt{\frac{\lambda^2}{4} + \frac{\lambda}{2M}}, \tag{8}$$

which, as expected, is a monotone increasing function of λ and a monotone decreasing function of M. To ensure the feasibility of the capacity management problem, we require that the overall daily service capacity C is not lower than A^*. Formally, we impose the following assumption, which will be required throughout our paper:

$$C \ge A^* = \frac{\lambda}{2} + \sqrt{\frac{\lambda^2}{4} + \frac{\lambda}{2M}}. \tag{9}$$

We conclude this subsection by stating a connection between the assumption (9) and the service level constraint (7).

LEMMA 1. *Consider the modified $M/D/1$ queue. For any $\theta \in [0,1]$ and any $Z \ge 0$, the service level constraint (7) is satisfied for any $A \in [A^*, C]$.*

3.3. Structure of Contract Payments and Contracting Process

We assume that both the purchaser and the provider are risk-neutral. In particular, for the provider of type t, the expected profit is obtained by combining the transfer payment (14) with the patient-diverting and overtime costs, (2) and (1):

$$\begin{aligned}\Pi_a^t\left(A^t, Z^t\right) &= T^t\left(A^t, Z^t\right) - O^t(A^t, Z^t) - P(A^t, Z^t) \\ &= T^t(A^t, Z^t) - o^t E_{D_0}[(D_0 - C + A^t)^+] - b\lambda(1-\theta)Pr(X(A^t, Z^t) \geq Z^t). \qquad (10)\end{aligned}$$

Similarly, the purchaser minimizes the expected cost

$$\Pi_p = pT^H + (1-p)T^L, \qquad (11)$$

while ensuring that the patient waiting-time target (7) is met.

A performance-based contract, a special case of the general contract defined above, consists of two types of payments: an activity-based, FFS payment from the service purchaser to the provider, and the penalty payment that the purchaser extracts from the provider based on achieved performance. Specifically, a contract (r^t, l^t) designed for a provider of type t includes payment r^t paid to the provider for serving each patient and daily penalty l^t incurred by the provider for every day patients spend, on average, waiting for appointments. A clear advantage of such a contract is its simple form: the performance-based penalty term is *linear* in the expected patient waiting time. At the same time, more general *non-linear* contracts may allow the purchaser to design a better calibrated incentive structure. In Section 6, we will consider one such contract under which a performance-based penalty is imposed once the expected patient waiting time exceeds a critical value.

We assume that the FFS payment is the same for both advance and same-day patients. As the expected number of patients treated each day is equal to

$$\lambda_0 + \lambda(1 - (1-\theta)Pr(X(A^t, Z^t) \geq Z^t)), \qquad (12)$$

the expected daily FFS payment is

$$r^t\left(\lambda_0 + \lambda(1 - (1-\theta)Pr(X(A^t, Z^t) \geq Z^t))\right). \qquad (13)$$

On the penalty side, the expected daily amount is $l^t W_q(A^t, Z^t)/A^t$, so that the total expected daily transfer payment from the purchaser to the provider is given by

$$T^t\left(r^t, l^t, A^t, Z^t\right) = r^t\left(\lambda_0 + \lambda(1 - (1-\theta)Pr(X(A^t, Z^t) \geq Z^t)\right) - l^t \frac{W_q(A^t, Z^t)}{A^t}. \qquad (14)$$

In the healthcare economics literature, it is often assumed that the service provider is altruistic and derives additional, non-monetary utility from providing a service to patients (see, e.g., Kaarboe and Siciliani (2011)). In practice, however, it is very hard to evaluate such a utility contribution, and, thus, we limit our analysis to the provider that maximizes expected profit. It is interesting to note that even in the UK, where healthcare providers are publicly funded non-profit organizations, they are faced with increasing financial pressure on their incomes and budgets, and with increasing freedom to manage their assets and provide services. For example, NHS Foundation Trusts, despite being public service providers, enjoy a significant level of autonomy over their affairs and substantial financial flexibility, and are often described as profit maximizers (De Fraja (2000) and Miraldo *et al.* (2011)).

In our analysis, we focus on the structure of the general contract and the performance-based contract under different information settings, starting with the benchmark case of symmetric information, under which the provider's cost structure is known to the purchaser, and following up with the asymmetric information case in which the provider's cost information is private. The sequence of events during the contracting process is as follows. Under the symmetric information setting, the provider's type t (H or L) is revealed, and the purchaser sets the contract terms for the provider. Under asymmetric information, the purchaser determines the contract terms for each provider type and offers a menu consisting of two contracts to the provider. Next, the provider either accepts the offered contract (under the symmetric information setting) or selects one contract from the offered menu (under the asymmetric information setting) and delivers the contracted service. Finally, the total number of activities (served patients) is counted and the service performance (expected waiting time) is evaluated, after which the provider receives contractual compensation.

4. Symmetric Information Setting

Under the symmetric information setting, the purchaser learns the provider's type t before deciding on the contract terms, and can therefore formulate a contract tailored to the specific provider type. In some healthcare systems, the purchaser can also observe and verify the provider's capacity allocation decisions, and can make those decisions a part of the provider's contractual obligations. For example, in a number of European countries, an active use of centralized appointment and record keeping systems provides purchasing agencies with a visibility of providers' capacity management actions. In more decentralized healthcare delivery environments, such as the one used in the US, capacity allocation policies often constitute the provider's "private actions", which remain unobservable to the purchaser. In such environments, the purchaser cannot include capacity management actions in a provider's contract and has to rely on financial levers to incentivize the provider to act on the purchaser's behalf. Below we analyze both of these environments.

4.1. Observable and Contractible Actions: The First-Best Solution and The Infinite-Horizon Heuristic

If the provider's capacity allocation policy (A,Z) is observable and contractible, the purchaser can always force the provider to act on the purchaser's behalf. In particular, given that the provider is of type $t\in\{H,L\}$, the purchaser solves the following problem:

$$\begin{aligned}\min_{T^t,A^t,Z^t}\quad & T^t\left(A^t,Z^t\right) && (15)\\ \text{s.t.}\quad & \left(A^t,Z^t\right)\in\mathcal{R}\left(M,C,\theta,\lambda\right) && (16)\\ & \Pi_a^t\left(T^t,A^t,Z^t\right)\geq 0, && (17)\\ & T^t\geq 0, && (18)\end{aligned}$$

where

$$\mathcal{R}\left(M,C,\theta,\lambda\right)=\{(A,Z)|\frac{W_q(A,Z)}{A}\leq M,\frac{\theta\lambda}{2}+\sqrt{\frac{\theta^2\lambda^2}{4}+\frac{\theta\lambda}{2M}}\leq A\leq C,Z\in\mathcal{N}\}. \qquad (19)$$

The objective for the purchaser is to minimize the expected payout T^t, $t\in\{H,L\}$. The first constraint, (16), specifies a service-level requirement stating that the expected number of days a patient spends waiting for her appointment does not exceed M, and that A^t cannot be below $\frac{\theta\lambda}{2}+\sqrt{\frac{\theta^2\lambda^2}{4}+\frac{\theta\lambda}{2M}}$, the value that guarantees that the expected waiting-time target is met even for $Z^t=0$. The second constraint, (17), is a participation requirement which stipulates that the provider's expected profit has to be non-negative. The following proposition describes the optimal solution for the complete information problem, also known as the first-best solution.

Proposition 2. *(a) For* $\frac{\theta\lambda}{2}+\sqrt{\frac{\theta^2\lambda^2}{4}+\frac{\theta\lambda}{2M}}\leq A\leq C$, *let*

$$Z_M(A)=\max_{Z\in\mathcal{N}}\left(Z|\frac{W_q(A,Z)}{A}\leq M\right). \qquad (20)$$

The family of optimal first-best contracts $(T^t_{\text{FB}},A^t_{\text{FB}},Z^t_{\text{FB}})$ *is characterized by:*

$$\begin{aligned}A^t_{\text{FB}} &= \underset{\frac{\theta\lambda}{2}+\sqrt{\frac{\theta^2\lambda^2}{4}+\frac{\theta\lambda}{2M}}\leq A^t\leq C}{\arg\min}\left(o^tE_{D_0}[(D_0-C+A^t)^+]+b\lambda(1-\theta)Pr\left(X\left(A^t,Z^t_M\left(A^t\right)\right)\geq Z^t_M\left(A^t\right)\right)\right),\\ Z^t_{\text{FB}} &= Z_M\left(A^t_{\text{FB}}\right), \qquad (21)\end{aligned}$$

and

$$T^t_{\text{FB}}=o^tE_{D_0}[(D_0-C+A^t_{\text{FB}})^+]+b\lambda(1-\theta)Pr(X(A^t_{\text{FB}},Z^t_{\text{FB}})\geq Z^t_{\text{FB}}). \qquad (22)$$

(b) The first-best capacity allocation decisions A^t_{FB} *and* Z^t_{FB} *are non-increasing functions of* o^t *and non-decreasing functions of* b.

(c) The linear performance-based contract can achieve the optimal first-best performance if and only if

$$\begin{aligned}r^t_{\text{FB}} &= \frac{o^tE_{D_0}[(D_0-C+A^t_{\text{FB}})^+]+b\lambda(1-\theta)Pr(X(A^t_{\text{FB}},Z^t_{\text{FB}})\geq Z^t_{\text{FB}})+l^t_{\text{FB}}W_q(A^t_{\text{FB}},Z^t_{\text{FB}})/A^t_{\text{FB}}}{\lambda_0+\lambda(1-(1-\theta)Pr\left(X(A^t_{\text{FB}},Z^t_{\text{FB}})\geq Z^t_{\text{FB}}\right))},\\ l^t_{\text{FB}} &\in \mathcal{R}^+,t\in\{H,L\}. \qquad (23)\end{aligned}$$

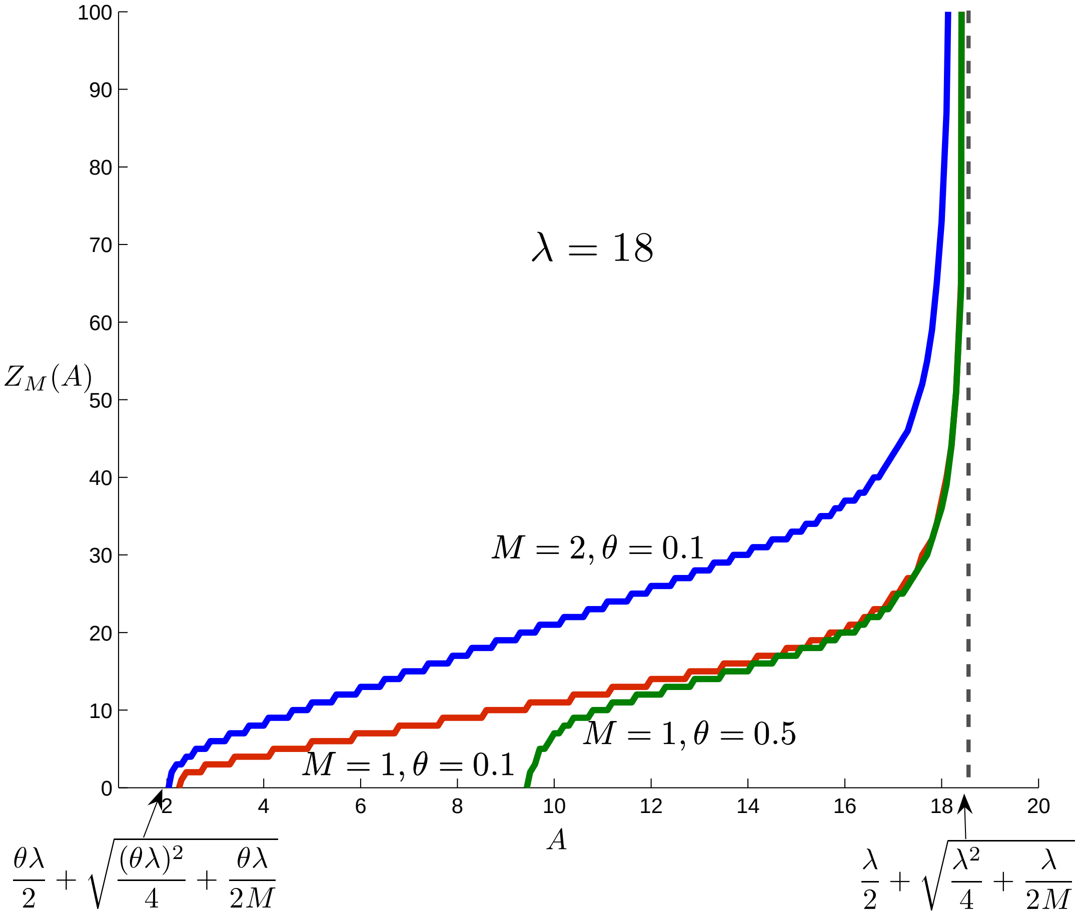

Figure 1: Matching appointment horizon $Z_M(A)$ as a function of the daily threshold A for different values of the expected waiting target M and the fraction of dedicated patients θ ($\lambda = 18$).

The results of Proposition 2 state that the first-best capacity allocation policy $(A^t_{\text{FB}}, Z^t_{\text{FB}})$ minimizes the sum of the expected overtime cost and the patient diverting cost, while ensuring, as (20) and (21) indicate, that the waiting-time constraint is satisfied in as tightly as possible. As follows from the monotonicity properties of $\frac{W_q(A,Z)}{A}$ described in Proposition 1, appointment horizon $Z_M(A)$ matching daily capacity for advance appointments, A, is an increasing function of A and M, and a decreasing function of θ, as illustrated in Figure 1. The first-best policy represents, as expected, a "centralized" capacity allocation solution, i.e., it maximizes the expected profit for the entire service supply chain. In addition, the optimal payment T^t is set to extract the entire surplus from the provider, so that $\Pi^t_a\left(T^t_{\text{FB}}, A^t_{\text{FB}}, Z^t_{\text{FB}}\right) = 0$. In a similar way, the optimal linear performance-based contract parameters, r^t_{FB} and l^t_{FB}, are set to extract the entire surplus from the provider, so that $\Pi^t_a\left(r^t_{\text{FB}}, l^t_{\text{FB}}, A^t_{\text{FB}}, Z^t_{\text{FB}}\right) = 0$. As (23) implies, there exists an infinite number of $(r^t_{\text{FB}}, l^t_{\text{FB}})$ pairs that achieve the first-best solution, so that the first-best contract can be cast in a performance-based

($l^t_{\text{FB}} > 0$) or a fee-for-service ($l^t_{\text{FB}} = 0$) format. It is important to note that the optimal value of the objective function for the first-best problem, $T^t_{\text{FB}}\,(r^t_{\text{FB}}, l^t_{\text{FB}}, A^t_{\text{FB}}, Z^t_{\text{FB}})$, does not depend on the choice of $(r^t_{\text{FB}}, l^t_{\text{FB}})$, but is rather determined by the capacity allocation policy $(A^t_{\text{FB}}, Z^t_{\text{FB}})$. In general, no closed-form expressions exist for Z^t_{FB} and A^t_{FB}, so the first-best capacity allocation policy has to be established numerically. However, sharper characterizations of the first-best controls are available for several special cases.

Corollary 1. *(a) For $o^t = 0, t \in \{H, L\}$, the first-best solution is given by*

$$\begin{aligned} A^t_{\text{FB}} &= C, \\ Z^t_{\text{FB}} &= \infty. \end{aligned} \tag{24}$$

Moreover, the optimal linear performance-based contract parameters are given by

$$\begin{aligned} r^t_{\text{FB}} &= \frac{\lambda l^t_{\text{FB}}}{2C\,(C-\lambda)\,(\lambda+\lambda_0)}, \\ l^t_{\text{FB}} &\in \mathcal{R}^+, t \in \{H, L\}. \end{aligned} \tag{25}$$

(b) For $b = 0$, the first-best solution is given by

$$\begin{aligned} A^t_{\text{FB}} &= \frac{\theta\lambda}{2} + \sqrt{\frac{\theta^2\lambda^2}{4} + \frac{\theta\lambda}{2M}}, \\ Z^t_{\text{FB}} &= 0. \end{aligned} \tag{26}$$

Moreover, the optimal linear performance-based contract parameters are given by

$$\begin{aligned} r^t_{\text{FB}} &= \frac{o^t E_{D_0}[(D_0 - C + A^t_{\text{FB}})^+] + l^t_{\text{FB}} M}{\lambda_0 + \theta\lambda}, \\ l^t_{\text{FB}} &\in \mathcal{R}^+, t \in \{H, L\}. \end{aligned} \tag{27}$$

(c) For $\theta = 1$, the first-best solution is given by

$$\begin{aligned} A^t_{\text{FB}} &= A^* = \frac{\lambda}{2} + \sqrt{\frac{\lambda^2}{4} + \frac{\lambda}{2M}}, \\ Z^t_{\text{FB}} &\in \mathcal{N}. \end{aligned} \tag{28}$$

Moreover, the optimal linear performance-based contract parameters are given by

$$\begin{aligned} r^t_{\text{FB}} &= \frac{o^t E_{D_0}[(D_0 - C + A^t_{\text{FB}})^+] + l^t_{\text{FB}} M}{\lambda_0 + \lambda}, \\ l^t_{\text{FB}} &\in \mathcal{R}^+, t \in \{H, L\}. \end{aligned} \tag{29}$$

Corollary 1 outlines the intuitive nature of the first-best capacity allocation policy: as the relative importance of the patient-diverting penalty cost over the overtime cost increases, the policy shifts from allocating the minimum feasible capacity to advance appointments while completely blocking flexible patients ($A=\frac{\theta\lambda}{2}+\sqrt{\frac{\theta^2\lambda^2}{4}+\frac{\theta\lambda}{2M}}$ and $Z=0$) to allocating the entire available capacity to advance appointments and serving the entire pools of dedicated and flexible patients ($A=C$ and $Z=\infty$). Note that for the provider serving only dedicated patients, the optimal capacity allocated to advance appointments, A^t_{FB}, does not depend on the provider's type. Thus, the expected cost to the purchaser of enforcing the waiting-time target, $T_{\mathrm{FB}}=pT^H_{\mathrm{FB}}+(1-p)T^L_{\mathrm{FB}}$, can be expressed as

$$T_{\mathrm{FB}}=pT^H_{\mathrm{FB}}+(1-p)T^L_{\mathrm{FB}}=(po^H+(1-p)o^L)E_{D_0}[(D_0-C+A^*)^+]. \tag{30}$$

In the "dedicated only" setting, both A^t_{FB} and the optimal expected payout value T^t_{FB} for either provider type are increasing in the average daily demand for advance appointments, λ, and decreasing in value for the waiting-time target, M. The latter property is illustrated in Figure 2 for the case of Poisson same-day demand with the rate $\lambda_0=2$, $\lambda=18$ and $C=20$. Note that the cost that the purchaser has to incur to ensure the patient waiting-time target of M weeks increases dramatically as M (expressed in days) becomes comparable to the expected demand rate for advance appointments λ. Both components of the capacity allocation policy, A^t_{FB} and Z^t_{FB}, play their own important roles in ensuring that the waiting-time target is achieved at the lowest cost. In particular, A^t_{FB} controls the overtime costs by reserving $C-A^t_{\mathrm{FB}}$ daily service slots for the use of same-day patients, while Z^t_{FB} serves as an important lever regulating flexible patient access to the service capacity. Figure 3 illustrates the first-best capacity allocation decisions as functions of the waiting-time target M in 9 settings characterized by different compositions of patient population ($\theta=0.1$, corresponding to mostly flexible patients, $\theta=0.5$, corresponding to an equal mix of dedicated and flexible patients, and $\theta=0.9$, corresponding to mostly dedicated patients) and different values of the ratio of overtime and patient diverting costs $o^t/b=1$, 5, and 10. We observe that in the setting where the cost of patient diversions is comparable to that of overtime service, A^t_{FB} remains relatively insensitive to the composition of the patient population or the service access requirements, and the first-best policy adjusts the allocation of service capacity almost entirely through changes in Z^t_{FB} that conform to the monotonicity properties illustrated in Figure 1: the higher is the fraction of flexible patients and the tolerance for service delays, the higher is the first-best appointment horizon. On the other hand, as financial penalties associated with patient diversions diminish, the composition of the patient population plays an increasingly important role

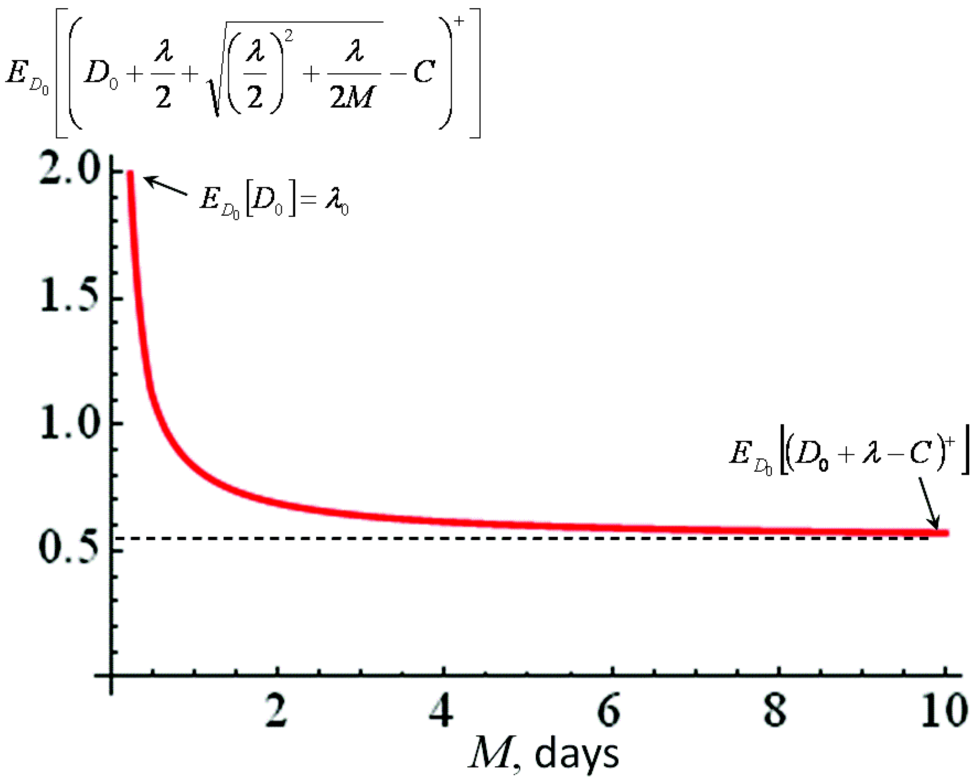

Figure 2: Optimal first-best transfer payment $T_{\text{FB}}/(po^H+(1-p)o^L)$ as a function of the waiting time target M for $\theta=1$, $\lambda=18$, Poisson same-day demand with rate $\lambda_0=2$, and $C=20$.

in shaping the first-best capacity allocation policy: while being largely insensitive to service level M, both A^t_{FB} and Z^t_{FB} change in a non-monotone fashion as functions of θ.

Our interactions with UK hospital managers point to an interesting observation: while the role of A^t_{FB} is well understood by practitioners, there exists a certain degree of reluctance about using Z^t_{FB} to protect the capacity for the use of dedicated patients. This reluctance is manifested in a policy under which the entire appointment horizon is made available through the CaB system: in our model, this corresponds to the setting $Z^t=\infty$. As Proposition 2 and Corollary 1 indicate, such an approach may be advisable in settings where the patient-diverting cost is high. Based on this approach, we can formally define a heuristic capacity allocation policy (which we call *infinite-horizon*, or IH, heuristic). In particular, under the IH heuristic, $Z^t_{\text{IH}}=\infty$ and $A^t_{\text{IH}}=A^*=\frac{\lambda}{2}+\sqrt{\frac{\lambda 2}{4}+\frac{\lambda}{2M}}$. Note that the IH capacity allocation policy ensures that the waiting-time target is met: $\frac{W_q(A^t_{\text{IH}},Z^t_{\text{IH}})}{A^t_{\text{IH}}}=M$. The following result establishes a bound on the performance of the IH heuristic in a general setting.

Lemma 2.

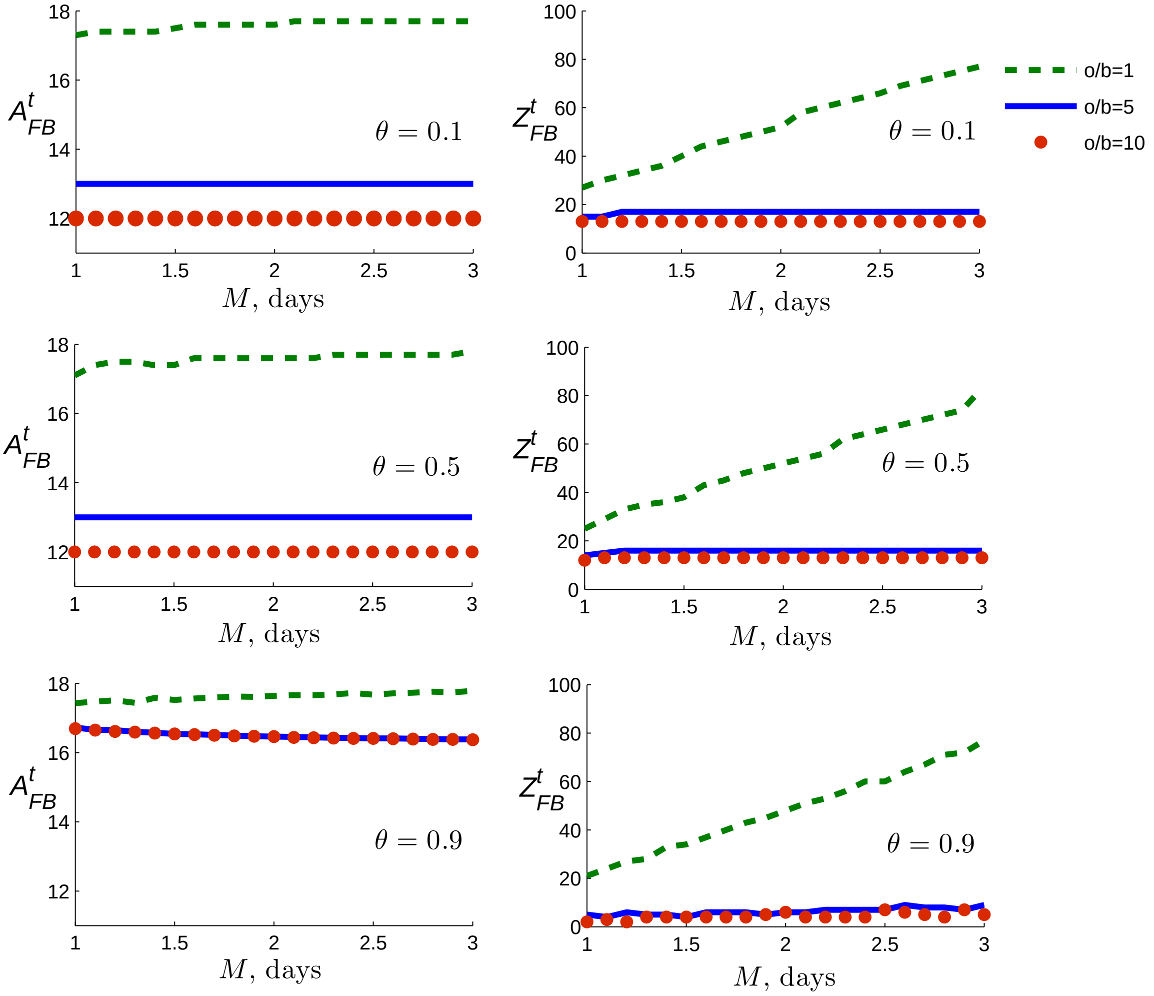

Figure 3: Optimal first-best capacity allocation decisions A^t_{FB} and Z^t_{FB} as functions of the waiting time target M for $\lambda = 18$, $\lambda_0 = 5$, and $C = 20$.

$$\frac{T_{\text{IH}}}{T_{\text{FB}}} \leq \frac{E_{D_0}\left[\left(D_0 - C + \frac{\lambda}{2} + \sqrt{\frac{\lambda^2}{4} + \frac{\lambda}{2M}}\right)^+\right]}{E_{D_0}\left[\left(D_0 - C + \frac{\theta\lambda}{2} + \sqrt{\frac{\theta^2\lambda^2}{4} + \frac{\theta\lambda}{2M}}\right)^+\right]}. \tag{31}$$

As shown by the result of Lemma 2, the IH heuristic is particularly effective in settings where the majority of patients are dedicated - an intuitive result in light of Proposition 2 and Corollary 1. On the other hand, as the fraction of flexible patients increases, a capacity allocation policy that ignores the necessity of reserving service capacity for dedicated patients fares increasingly poorly.

4.2. Private Actions: Implementing the First-Best Outcome under Information Symmetry

In our analysis above, we have assumed that the provider's capacity allocation decisions A and Z are both observable and contractible by the purchaser. In practice, however, observing and verifying the provider's decisions may be too difficult and/or too costly for the purchaser. In such "private action" settings, the contract terms offered by the principal do not include a specification of capacity allocation policies, and the provider will choose such policies to maximize his expected profit. Thus, to implement the first-best solution, the contract terms must be designed to induce the provider of type t to choose A^t and Z^t as his optimal decisions. Below, we consider three types of contracts that have been used in the past or are being used at present by the UK's National Health Service: the fixed lump-sum payment (block contract), the fee-for-service payment (FFS), and the linear performance-based contract (PBC).

Under the block contract, let T^t be the fixed lump-sum payment paid by the purchaser to the provider irrespective of the actual volume of provided services or the achieved service access level. The type-t provider's problem is

$$\max_{A^t,Z^t} \left(T^t - o^t E_{D_0}[(D_0 - C + A^t)^+] - b\lambda(1-\theta)Pr(X(A^t,Z^t) \geq Z^t)\right) \tag{32}$$

$$\text{s.t. } \theta\lambda \leq A^t \leq C, Z^t \in \mathcal{N}. \tag{33}$$

It is clear that the provider's optimal capacity allocation in this case is not influenced by T^t. In particular, the provider will choose $Z^t = \infty$ to eliminate the patient diversion cost. Also, since the overtime cost is decreasing in A^t, the provider will choose it to be at its lower-bound value $\theta\lambda$, resulting in a violation of the purchaser's waiting-time constraint. Thus, under the private action setting, the block contract cannot achieve the first-best outcome.

Under the FFS contract, the purchaser controls only the payment amount r^t for each patient served by the provider, producing a standard linear price contract (Bolton and Dewatripont (2005)). Under the FFS approach, the transfer payment can be written as

$$T^t = r^t\left(\lambda_0 + \lambda(1-(1-\theta)Pr(X(A^t,Z^t) \geq Z^t))\right). \tag{34}$$

For any r^t, the type-t provider's problem is

$$\max_{A^t,Z^t} \left(r^t\left(\lambda_0 + \lambda(1-(1-\theta)Pr(X(A^t,Z^t) \geq Z^t))\right) - o^t E_{D_0}[(D_0 - C + A^t)^+] - b\lambda(1-\theta)Pr(X(A^t,Z^t) \geq Z^t)\right)$$

$$\text{s.t. } \theta\lambda \leq A^t \leq C, Z^t \in \mathcal{N}. \tag{35}$$

Note that the provider's objective is increasing in Z^t. Thus, similar to the block contract, the provider will choose $Z^t = \infty$ and $A^t = \theta\lambda$, which will violate the purchaser's waiting-time constraint. Thus, the FFS contract cannot achieve the first-best outcome either.

Now, let us consider a linear performance-based contract under which the fee-for-service payment is adjusted by the performance penalty based on the achieved expected patient waiting time, so that the transfer payment is given by

$$T^t = r^t\left(\lambda_0 + \lambda(1-(1-\theta)Pr(X(A^t,Z^t) \geq Z^t))\right) - l^t \frac{W_q(A^t,Z^t)}{A^t}. \tag{36}$$

The following proposition identifies linear PBC contract parameters that achieve the first-best outcome for the setting in which all patients are dedicated.

Proposition 3. *For $\theta = 1$, the first-best outcome is obtained by using the contract*

$$\tilde{r}^t = \frac{o^t}{\lambda+\lambda_0}\left(\frac{\lambda\left(1-F_{D_0}\left(C-\frac{\lambda}{2}-\sqrt{\frac{\lambda^2}{4}+\frac{\lambda}{2M}}\right)\right)}{4M\sqrt{\frac{\lambda^2}{4}+\frac{\lambda}{2M}}} + E_{D_0}\left[\left(D_0 - C + \frac{\lambda}{2} + \sqrt{\frac{\lambda^2}{4}+\frac{\lambda}{2M}}\right)^+\right]\right), \tag{37}$$

$$\tilde{l}^t = o^t\left(\frac{1-F_{D_0}\left(C-\frac{\lambda}{2}-\sqrt{\frac{\lambda^2}{4}+\frac{\lambda}{2M}}\right)}{4M^2\sqrt{\frac{\lambda^2}{4}+\frac{\lambda}{2M}}}\right), \tag{38}$$

where $\lambda_0 = E_{D_0}[D_0]$.

Proposition 3 states that in the private-action setting, the performance-based contract parameters, r^t and l^t, are no longer arbitrarily selected from a set described in Corollary 1, but are uniquely determined by (37) and (38). Note that shorter waiting-time target values lead to higher activity-based price levels, higher performance-based penalties, and higher transfer payments:

Corollary 2. *For $\theta = 1$, the optimal contract terms, $\tilde{r}^t$ and $\tilde{l}^t$, as well as the resulting transfer payment, $\tilde{T}^t$, are monotone decreasing functions of M for any provider type t.*

In summary, our analysis of the private-action setting indicates that even when the purchaser possesses complete information about provider's cost structure, the fee-for-service contract alone cannot support the waiting-time target, and the performance-based incentive is required to ensure that the provider will allocate adequate capacity to serve advance appointments.

5. The Asymmetric Information Setting

The informational advantage of service providers over purchasers is expected to infuse inefficiency into service capacity allocation outcomes. In the analysis below, we explore the influence of information asymmetry regarding the value of the provider's overtime cost on the structure of the optimal service contracts. As in the case of information symmetry, we start by considering the case in which the provider's capacity allocation actions are both observable and contractible.

5.1. Observable and Contractible Actions: The Second-Best Solution

The information asymmetry in assessing the provider's overtime costs leads to the adverse selection problem. In particular, given a contract menu $\{T^t, A^t, Z^t\}$, $t \in \{H, L\}$, a provider of type H may choose to select a contract designed for a provider of type L, if the latter gives him a higher payoff. To deal with this possibility, the purchaser must design a contract menu by applying the revelation principle. More specifically, let Π_a^{ts}, $t, s \in \{H, L\}$ denote the expected payoff for the provider of type t who reports to be of type s (in other words, who chooses a contract designed for type s providers):

$$\Pi_a^{ts}(T^s, A^s, Z^s) = T^s\left(A^s, Z^s\right) - o^t E_{D_0}[(D_0 - C + A^s)^+] - b\lambda(1-\theta)Pr(X(A^s, Z^s) \geq Z^s). \tag{39}$$

Note that $\Pi_a^t\left(T^t, A^t, Z^t\right)$ defined in (10) is equivalent to $\Pi_a^{tt}(T^t, A^t, Z^t)$. The purchaser's problem can be formulated as follows:

$$\begin{aligned}
\min_{T^t, A^t, Z^t, t \in \{H,L\}} \quad & \left(pT^H\left(A^H, Z^H\right) + (1-p)T^L\left(A^L, Z^L\right)\right) & (40)\\
\text{s.t.} \quad & (A^t, Z^t) \in \mathcal{R}\left(M, C, \theta, \lambda\right), t \in \{H, L\} & (41)\\
& \Pi_a^{tt}(T^t, A^t, Z^t) \geq 0, t \in \{H, L\} & (42)\\
& \Pi_a^{tt}(T^t, A^t, Z^t) \geq \Pi_a^{ts}(T^s, A^s, Z^s),\ t, s \in \{H, L\}, s \neq t, & (43)\\
& T^t \geq 0, t \in \{H, L\}. & (44)
\end{aligned}$$

The waiting-time target and stability constraints (41) and the individual rationality constraints (42) are the analogues of the constraints (16) and (17) in the symmetric information setting. Constraint (43) ensures correct matching between provider types and contract types. The contract optimizing the purchaser's objective is usually labeled as the second-best solution. Note that in the case of a linear performance-based contract, $T^t(A^t, Z^t)$ takes the special form of

$$T^t(A^t, Z^t) = r^t\left(\lambda_0 + \lambda(1 - (1-\theta)Pr(X(A^t, Z^t) \geq Z^t)\right) - l^t \frac{W_q(A^t, Z^t)}{A^t}. \tag{45}$$

Proposition 4. *(a) The family of optimal second-best contracts is characterized by*

$$\begin{aligned}
A_{\text{SB}}^t &= \underset{\frac{\theta\lambda}{2} + \sqrt{\frac{\theta^2\lambda^2}{4} + \frac{\theta\lambda}{2M}} \leq A^t \leq C}{\arg\min} \left(\hat{o}^t E_{D_0}[(D_0 - C + A^t)^+] + b\lambda(1-\theta)Pr\left(X\left(A^t, Z_M^t\left(A^t\right)\right) \geq Z_M^t\left(A^t\right)\right)\right),\\
Z_{\text{SB}}^t &= Z_M\left(A_{\text{SB}}^t\right),
\end{aligned} \tag{46}$$

where

$$\begin{aligned}
\hat{o}^H &= o^H + \frac{1-p}{p}\left(o^H - o^L\right),\\
\hat{o}^L &= o^L,
\end{aligned} \tag{47}$$

and $Z_M\left(A\right)$ as defined in (20).

(b) The optimal values of the expected payments to providers are given by

$$\begin{aligned} T^H_{\text{SB}} &= o^H E_{D_0}[(D_0 - C + A^H_{\text{SB}})^+] + b\lambda(1-\theta)Pr(X(A^H_{\text{SB}}, Z^H_{\text{SB}}) \geq Z^H_{\text{SB}}),\\ T^L_{\text{SB}} &= o^L E_{D_0}[(D_0 - C + A^L_{\text{SB}})^+] + (o^H - o^L)E_{D_0}[(D_0 - C + A^H_{\text{SB}})^+]\\ &\quad + b\lambda(1-\theta)Pr(X(A^L_{\text{SB}}, Z^L_{\text{SB}}) \geq Z^L_{\text{SB}}). \end{aligned} \tag{48}$$

(c)

$$A^H_{\text{SB}} \leq A^H_{\text{FB}} \leq A^L_{\text{FB}} = A^L_{\text{SB}}, \tag{49}$$

and

$$Z^H_{\text{SB}} \leq Z^H_{\text{FB}} \leq Z^L_{\text{FB}} = Z^L_{\text{SB}}. \tag{50}$$

where A^t_{FB} and Z^t_{FB} are the first-best capacity allocation controls defined in (21).

(d) The linear performance-based contract can achieve the optimal second-best performance if and only if

$$\begin{aligned} r^H_{\text{SB}} &= \frac{o^H E_{D_0}[(D_0 - C + A^H_{\text{SB}})^+] + b\lambda(1-\theta)Pr(X(A^H_{\text{SB}}, Z^H_{\text{SB}}) \geq Z^H_{\text{SB}}) + l^H_{\text{SB}} W_q(A^H_{\text{SB}}, Z^H_{\text{SB}})/A^H_{\text{SB}}}{\lambda_0 + \lambda(1 - (1-\theta)Pr\,(X(A^H_{\text{SB}}, Z^H_{\text{SB}}) \geq Z^H_{\text{SB}}))},\\ r^L_{\text{SB}} &= \frac{o^L E_{D_0}[(D_0 - C + A^L_{\text{SB}})^+] + b\lambda(1-\theta)Pr(X(A^L_{\text{SB}}, Z^L_{\text{SB}}) \geq Z^L_{\text{SB}}) + l^L_{\text{SB}} W_q(A^L_{\text{SB}}, Z^L_{\text{SB}})/A^L_{\text{SB}}}{\lambda_0 + \lambda(1 - (1-\theta)Pr\,(X(A^L_{\text{SB}}, Z^L_{\text{SB}}) \geq Z^L_{\text{SB}})}\\ &+ \frac{(o^H - o^L)E_{D_0}[(D_0 - C + A^H_{\text{SB}})^+]}{\lambda_0 + \lambda(1 - (1-\theta)Pr\,(X(A^L_{\text{SB}}, Z^L_{\text{SB}}) \geq Z^L_{\text{SB}})},\\ l^t_{\text{SB}} &\in \mathcal{R}^+, t \in \{H, L\}. \end{aligned} \tag{51}$$

As Proposition 4 states, the second-best capacity allocation policy is obtained by solving two separate optimization problems, one for each provider type, with a structure identical to that of the first-best problem (21). In particular, the optimization problem for the low-cost provider is completely identical to (21), while that for the high-cost provider uses the value of the overtime cost for this provider type adjusted upward due to the presence of information asymmetry. Consequently, while the second-best capacity allocation policy intended for the low-cost provider replicates the first-best policy for this provider type, the second-best capacity allocation policy intended for the high-cost provider differs from the corresponding first-best solution. In particular, the daily capacity allocated to advance appointments under the second-best solution (A^H_{SB}) is, in general, lower than that under the first-best solution (A^H_{FB}) - inefficiency created by information asymmetry. The transfer payment structure is also different from that of the first-best solution. Although the payout to the high-cost provider is still equal to its operational cost (equal to the sum of the overtime cost and the patient diverting cost), the payout to the low-cost provider is higher than its operational cost: an additional information rent, $(o^H - o^L)E_{D_0}[(D_0 - C + A^H_{\text{SB}})^+]$, is paid to the low-cost provider as a result of the existing information asymmetry. Consequently, the high-cost provider's net surplus is zero, just as in the symmetric information setting, while the low-cost provider's net surplus is equal to the information rent: a payout structure similar to one outlined in the health economics literature (De Fraja (2000)).

Comparing the capacity allocation policies for the low- and the high-cost providers indicates that, in both the first-best and the second-best solutions, the high-cost provider tends to release less capacity to advance appointments and to have a shorter appointment horizon. Recall that the effective daily demand for appointments is $\lambda(1-(1-\theta)Pr(X(A,Z)\geq Z))$. The results of Proposition 4 (c) and Proposition 1 (b) imply that

$$\lambda(1-(1-\theta)Pr(X(A^H_{\text{SB}},Z^H_{\text{SB}})\geq Z^H_{\text{SB}}))\leq\lambda(1-(1-\theta)Pr(X(A^L_{\text{SB}},Z^L_{\text{SB}})\geq Z^L_{\text{SB}})), \tag{52}$$

which indicates that a provider of type L serves more patients than a provider of type H (while not necessarily receiving a higher total transfer payment or a higher transfer payment per appointment).

Closed-form expressions for the second-best contract parameters can be obtained in the same special cases described in Corollary 1. In particular, in all of these special cases ($o^t=0, t\in\{H,L\}$, or $b=0$, or $\theta=1$) the second-best and the first-best capacity allocation parameters intended for the high-cost provider coincide, and so do the second-best and the first-best solutions.

Similar to the first-best solution, if both A^t and Z^t are observable and contractible, the second-best outcome can be achieved by either block, FFS contracts, or linear PBCs because the optimal payout to each provider type is completely determined by the capacity allocation policy. In the previous section, we have also shown that in the absence of such information asymmetry, neither block nor FFS contracts can reproduce the first-best solution in settings with private provider actions. At the same time, there exists a unique linear PBC that achieves this task when $\theta=1$, as specified by Proposition 3. However, as shown below, if private provider actions are allowed in the asymmetric information setting, the linear PBC is no longer able to achieve the second-best solution even in the setting where all patients are dedicated.

5.2. Performance-Based Contracts under Information Asymmetry with Private Actions

In this section we establish that, in general, the second-best outcome cannot be implemented under information asymmetry using a linear PBC. In particular, we focus on the special case of a hospital serving only dedicated patients ($\theta=1$). Note that in this case the value of Z^t does not influence appointment dynamics or cost structure, and, therefore, the capacity allocation policy reduces to choosing the daily appointment threshold level A^t. Given a menu of linear PBCs, $(r^t, l^t), t \in \{H, L\}$, the type-$t$ provider who reports to be of type s solves the following optimization problem:

$$\max_{\theta\lambda \le A^{ts} \le C} \left(\Pi_a^{ts}(r^s, l^s, A^{ts}) \equiv T^s(r^s, l^s, A^{ts}) - o^t E_{D_0}[(D_0 - C + A^{ts})^+]\right), s \in \{H, L\}, \tag{53}$$

where $T^s(r^s, l^s, A^{ts}) = r^s(\lambda + \lambda_0) - \frac{l^s \lambda}{2A^{ts}(A^{ts} - \lambda)}$ is the transfer payment to the provider of type t who reports to be of type s. Denote the solution of the above optimization problem by A_{PA}^{ts}. The following proposition provides a partial characterization of the provider's optimal capacity allocation decision.

Proposition 5. *Let $\theta = 1$. Then, for any menu of contracts $(r^t, l^t), t \in \{H, L\}$,*

(a) $A_{\text{PA}}^{LH} \ge A_{\text{PA}}^{HH}$. In particular, if $\lambda < A_{\text{PA}}^{HH} < C$, then $A_{\text{PA}}^{LH} > A_{\text{PA}}^{HH}$.

(b) $A_{\text{PA}}^{LL} \ge A_{\text{PA}}^{HL}$. In particular, if $\lambda < A_{\text{PA}}^{HL} < C$, then $A_{\text{PA}}^{LL} > A_{\text{PA}}^{HL}$.

(c) A_{PA}^{ts} is increasing in l^s. In particular, $A_{\text{PA}}^{ts} \le A^* = \frac{\lambda}{2} + \sqrt{\frac{\lambda^2}{4} + \frac{\lambda}{2M}}$ if and only if $l^s \le \tilde{l}^t$, where

$$\tilde{l}^t = o^t \left(\frac{1 - F_{D_0}\left(C - \frac{\lambda}{2} - \sqrt{\frac{\lambda^2}{4} + \frac{\lambda}{2M}}\right)}{4M^2 \sqrt{\frac{\lambda^2}{4} + \frac{\lambda}{2M}}} \right), s, t \in \{H, L\}. \tag{54}$$

Proposition 5 shows that given any menu of contracts, the high-cost provider cannot choose higher capacity level than the low-cost provider. In addition, it states that the provider's capacity allocated to advance appointments, as expected, is increasing in the waiting-time penalty cost. In particular, when this penalty cost is low enough, the provider chooses a capacity level below A^*, which, in turn, violates the patient waiting-time requirement.

In the private action setting, the purchaser of outpatient services solves the following optimization problem:

$$\min_{r^t, l^t, t \in \{H, L\}} \quad \left(pT^H(r^H, l^H, A_{\text{PA}}^{HH}) + (1-p)T^L(r^L, l^L, A_{\text{PA}}^{LL})\right) \tag{55}$$

$$\text{s.t.} \quad A_{\text{PA}}^{HH} \ge A^*, \tag{56}$$

$$A_{\text{PA}}^{LL} \ge A^*, \tag{57}$$

$$\Pi_a^{HH}\left(r^H, l^H, A_{\text{PA}}^{HH}\right) \ge 0, \tag{58}$$

$$\Pi_a^{LL}\left(r^L, l^L, A_{\text{PA}}^{LL}\right) \ge 0, \tag{59}$$

$$\Pi_a^{HH}\left(r^H, l^H, A_{\text{PA}}^{HH}\right) \ge \Pi_a^{HL}\left(r^L, l^L, A_{\text{PA}}^{HL}\right), \tag{60}$$

$$\Pi_a^{LL}\left(r^L, l^L, A_{\text{PA}}^{LL}\right) \ge \Pi_a^{LH}\left(r^H, l^H, A_{\text{PA}}^{LH}\right), \tag{61}$$

$$r^t \ge 0, l^t \ge 0, t \in \{H, L\}. \tag{62}$$

While it is impossible to obtain closed-form expressions for the optimal contract parameters in (55)-(62), a partial characterization is provided below.

PROPOSITION 6. *When $\theta = 1$, the optimal contract, $(r^t_{\text{PA}}, l^t_{\text{PA}})\,, t \in \{H, L\}$, can be characterized as follows:*

(a) $l^t_{\text{PA}} \geq \tilde{l}^t$, $t \in \{H, L\}$.

(b) $A^{LL}_{\text{PA}} > A^*$.

(c) $\Pi^{HH}_a(r^H, l^H, A^{HH}_{\text{PA}}) = 0, \Pi^{LL}_a(r^L, l^L, A^{LL}_{\text{PA}}) > (o^H - o^L)E_{D_0}[(D_0 - C + A^*)^+]$.

(d) $pT^H(r^H, l^H, A^{HH}_{\text{PA}}) + (1-p)T^L(r^L, l^L, A^{LL}_{\text{PA}}) > o^H E_{D_0}[(D_0 - C + A^*)^+]$.

Proposition 6 implies that compared to the first-best capacity allocation, A^*, providers of both types tend to allocate higher capacities to serving advance appointments. In particular, the low-cost provider's capacity allocation is strictly larger than A^*, indicating that information asymmetry distorts the provider's behavior, which results in a loss of efficiency. Parts (c) and (d) of Proposition 6 provide further evidence of PBC's inability to achieve the second-best outcome: the purchaser provides a higher rent to the low-cost provider, and, overall, pays more for the same level of service than she does in the second-best solution.

6. Threshold-Penalty Performance-Based Contracts

Proposition 3 shows that the linear PBC can achieve the first-best performance when $\theta = 1$ (though not necessarily for an arbitrary $\theta < 1$). Proposition 6 shows that even when $\theta = 1$, the linear PBC cannot achieve the second-best performance and cannot coordinate the service supply chain. Below we show that these shortcomings can be remedied if one extends the analysis to include contracts with non-linear penalties for patient wait times. In particular, we focus on a simple threshold-penalty contract structure, under which a) the provider receives a fixed payment F, and b) a fixed penalty K is imposed on a provider if and only if the waiting-time target is not achieved. In our analysis we use the notation (F, K) to designate such a contract.

The following result describes a family of (F, K) contracts that achieve the first-best performance for any composition of patient population.

PROPOSITION 7. *Consider the symmetric information setting with private actions and let*

$$F^t = o^t E_{D_0}\left[(D_0 - C + A^t_{\text{FB}})^+\right] + b\lambda(1-\theta)Pr(X(A^t_{\text{FB}}, Z^t_{\text{FB}}) \geq Z^t_{\text{FB}}), \tag{63}$$

and let K be the positive constant such that

$$K > F^t - o^t E_{D_0}[(D_0 - C + \theta\lambda)^+]. \tag{64}$$

Consider a threshold-penalty contract under which a provider of type t receives a payment of F^t if the waiting-time constraint is satisfied and a payment of $F^t - K$ if it isn't:

$$T^t = \begin{cases} F^t, & \text{if } W_q(A^t, Z^t)/A^t \leq M, \\ F^t - K, & \text{if } W_q(A^t, Z^t)/A^t > M. \end{cases} \tag{65}$$

Any threshold-penalty performance-based contract specified by (63)-(65) achieves the first-best outcome.

In the asymmetric information setting, such a threshold-penalty contract structure can achieve the second-best performance in the case of dedicated-only patients.

Proposition 8. *Consider the asymmetric information setting with private actions and a threshold-penalty PBC $(\underline{F}, \underline{K})$ defined by*

$$\underline{F} = F^H, \tag{66}$$

and

$$\underline{K} = F^H - o^L E_{D_0}[(D_0 - C + \theta\lambda)^+], \tag{67}$$

where F^H is given by (63).

(a) Contract $(\underline{F}, \underline{K})$ minimizes the expected provider's cost among all threshold-penalty PBCs.

(b) Under the contract $(\underline{F}, \underline{K})$, the optimal capacity allocation policy for each provider type is described by

$$\begin{aligned} A^H_{\text{TP}}(\underline{F}, \underline{K}) &= A^H_{\text{FB}}, \\ Z^H_{\text{TP}}(\underline{F}, \underline{K}) &= Z^H_{\text{FB}}, \\ A^L_{\text{TP}}(\underline{F}, \underline{K}) &= A^L_{\text{FB}}, \\ Z^L_{\text{TP}}(\underline{F}, \underline{K}) &= Z^L_{\text{FB}}, \end{aligned} \tag{68}$$

and the resulting expected transfer payments are

$$T^H(\underline{F}, \underline{K}) = T^L(\underline{F}, \underline{K}) = \underline{F}. \tag{69}$$

(c) The threshold-penalty PBC $(\underline{F}, \underline{K})$ achieves the second-best solution for $\theta = 1$:

$$\begin{aligned} A^H_{\text{TP}}(\underline{F}, \underline{K}) &= A^L_{\text{TP}}(\underline{F}, \underline{K}) = A^*, \\ T^H(\underline{F}, \underline{K}) &= T^L(\underline{F}, \underline{K}) = o^H E_{D_0}[(D_0 - C + A^*)^+]. \end{aligned} \tag{70}$$

One important advantage of the $(\underline{F}, \underline{K})$ contract is its relative simplicity as, instead of a menu of contracts, it offers the same terms to both provider types. Note that in the case of $\theta = 1$, the second-best solution coincides with the first-best one, and, therefore, the threshold-penalty contract coordinates the system.

As the results of Proposition 2 (c) indicate, in settings with mixed patient populations the threshold-penalty PBC may no longer be able to achieve the second-best performance. Figure 4 depicts the second-best solution (A_{SB}^H, Z_{SB}^H, A_{SB}^L, Z_{SB}^L) and the threshold-penalty PBC solution (A_{TP}^H, Z_{TP}^H, A_{TP}^L, Z_{TP}^L) in the same patient-mix settings as in Figure 3: $\theta = 0.1$ (mostly flexible patients), $\theta = 0.5$ (an equal mix of dedicated and flexible patients), and $\theta = 0.9$ (mostly dedicated patients). The capacity allocation policies shown in Figure 4 prompt several observations. First, high-cost providers never allocate more capacity for advance appointments, either in terms of the number of daily appointments or in terms of the appointment horizon, than do low-cost providers. Second, the optimal allocation policies for the low-cost providers under the $(\underline{F}, \underline{K})$ PBC and in the second-best solution coincide. Third, under the $(\underline{F}, \underline{K})$ PBC contract, both the daily appointment capacity and the appointment horizon selected by high-cost providers are always between the corresponding allocations in the second-best solution for the high-cost providers and the corresponding allocations in the second-best solution for the low-cost providers. Thus, the threshold-penalty PBC does not always achieve the second-best solution, and the corresponding loss of efficiency occurs through the capacity allocation policies of the high-cost providers. Finally, consistent with the result of part (c) of Proposition 8, the efficiency gap reduces as the patient population mix shifts towards mostly dedicated patients.

7. Conclusions

As an ever increasing number of healthcare organizations recognize service access as an important component of the quality of healthcare services, performance-based contracts (PBC) that include access performance measures gain equally increasing popularity. In our paper we study an approach to contracting for outpatient services used in the UK under the aegis of the National Health Service. Two features of this approach are of particular importance for our analysis: an online system ("Choose-and-Book") for managing advance appointments, and explicit penalties imposed by purchasers on providers for delaying patient services. Faced with contracts that include compensation for provided services as well as penalties for denying or delaying service, hospitals and individual physicians respond with a policy for allocating their limited service capacity between urgent and non-urgent patients, with the latter group comprised of dedicated patients who prefer to receive service from the medical facility of their choice even if the wait involved is longer, and flexible patients who will prefer another provider in order to shorten their wait for an appointment. By designing a performance-based contract, a purchaser of healthcare services aims to achieve a particular service access goal (expressed in terms of patient appointment waiting time) at the lowest possible cost. In practice, this task is often complicated by information asymmetry between the provider and the purchaser of services. For such a setting, we derive the properties of the first-best

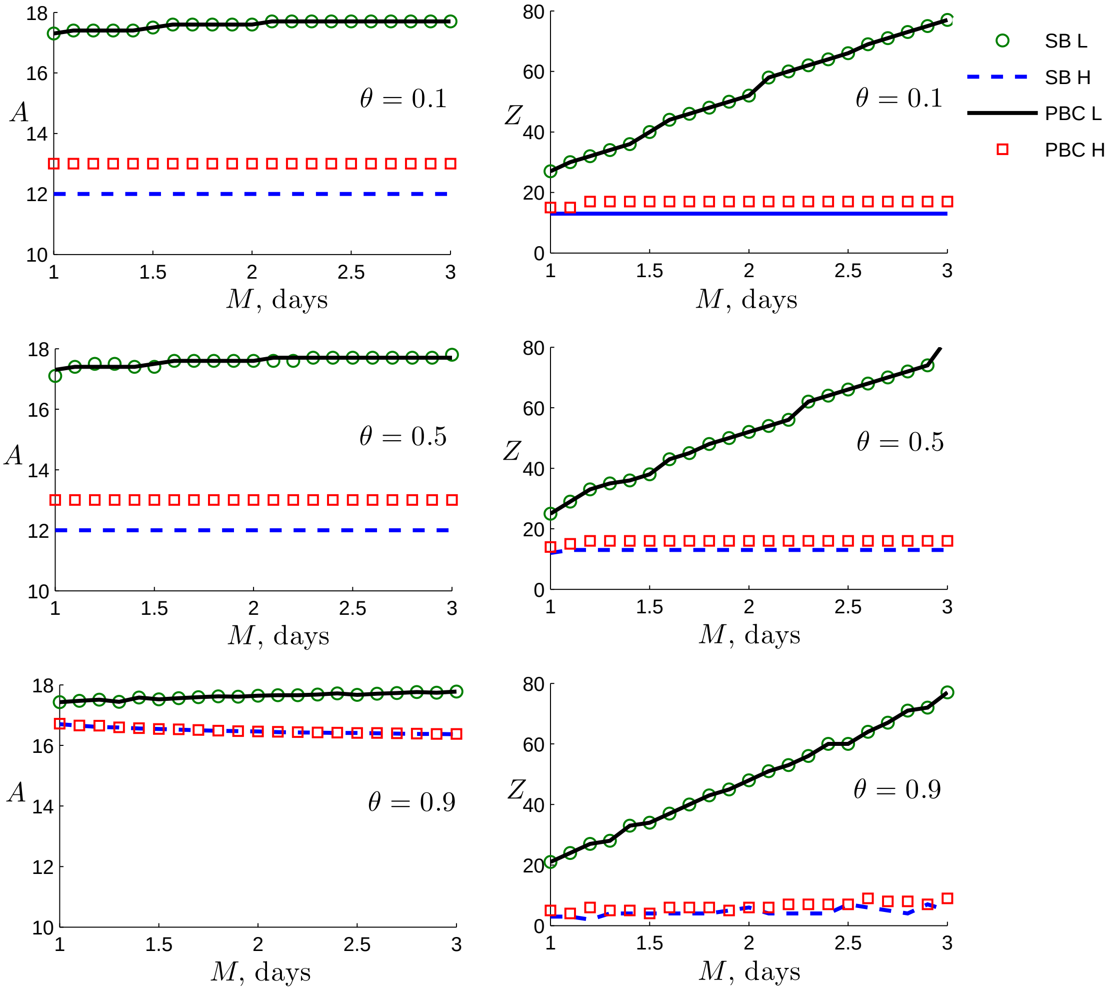

Figure 4: The optimal second-best solution $(A^H_{\text{SB}}, Z^H_{\text{SB}}, A^L_{\text{SB}}, Z^L_{\text{SB}})$ **and the optimal solution** $(A^H_{\text{TP}}, Z^H_{\text{TP}}, A^L_{\text{TP}}, Z^L_{\text{TP}})$ **for the threshold-penalty PBC as functions of the waiting-time target** M **for different values of** θ **(0.1, 0.5, 0.9),** $\lambda = 18$**, Poisson same-day demand with rate** $\lambda_0 = 5$**,** $o^H/b = 5$**,** $o^L/b = 1$**,** $p = 0.5$ **and** $C = 20$**. The top, middle and bottom subplots are for the settings of** $\theta = 0.1, 0.5, 0.9$**, respectively.**

and the second-best solutions for different patient population mixes by modeling the appointment dynamics as that of an $M/D/1$ queueing system. We show, in particular, that a linear PBC is guaranteed to achieve coordination only in the case of dedicated-only patients, and that it fails to achieve the second-best outcomes. As a remedy, we propose a simple threshold-penalty contract that always achieves the first-best performance and that also produces the second-best outcome in the case of dedicated-only patients.

An important feature of real-life capacity allocation decisions made by care providers is their multi-dimensional and dynamic nature. In the present work, we have adopted a simplifying approach to modeling these decisions by assuming a two-dimensional, open-loop provider's response that has allowed us to focus on important contractual issues while capturing important capacity allocation trade-offs. We believe future research can build on our findings by incorporating more complex and more realistic features of day-to-day appointment accumulation and service dynamics. On the contract design side, more investigation is needed into the nature of non-linear penalty contracts that can close the information-asymmetry-generated efficiency gap for an arbitrary patient mix. This line of research is particularly important in view of the increasing complexity of emerging performance-based contract structures (NHS Contract (2008)).

Acknowledgment

We are grateful to many professionals at Addenbrooke's Hospital in Cambridge, UK, and at the UK's National Health System for their generous assistance in helping us understand the functioning mechanisms of the Choose-and-Book online appointment management system, as well as the structure of Payment-by-Results contracts.

References

Arrow, K. J. 1963. Uncertainty and the welfare economics of medical care. *American Economic Review* **53**(5) 941–973.

Bloom, G., H. Standing, and R. Lloyd. 2008. Markets, information asymmetry and health care: Towards new social contracts. *Social Science and Medicine* **66**(10) 2076–2087.

Bolton, P., and M. Dewatripont. 2005. *Contract Theory.* MIT Press.

Brun, O., and J. Garcia. 2000. Analytical solution of finite capacity $M/D/1$ queues. *Journal of Applied Probability* **37**(4) 1092–1098.

Cachon, G.P. 2003. Supply chain coordination with contracts. *Handbooks in Operations Research and Management Science: Supply Chain Management.* Edited by S. Graves and T. de Kok. North Holland.

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

Farrar, S., J. Sussex, D. Yi, M. Sutton, M. Chalkley, T. Scott, and A. Ma. 2007. *National Evaluation of Payment by Results: Report to the Department of Health.* Health Economics Research Unit (HERU).

Fisher, E.S., D.E. Wennberg, T.A. Stukel, and D.J. Gottlieb. 2004. Variations in the longitudinal efficiency of academic medical centers. *Health Affairs*, Suppl. Web. Exclusive VAR19–32.

Fuloria, P.C., and S.A. Zenios. 2001. Outcomes-Adjusted Reimbursement in a Health-Care Delivery System. *Management Science*, **47**(6) 735–751.

Garnett, O., A. Mandelbaum, and M. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing and Service Operations Management* **4**(3) 208–227.

Goddard, M., R. Mannion, and P. Smith. 2000. Enhancing performance in health care: A theoretical perspective on agency and the role of information. *Health Economics* **9** 95–107.

Gupta, D., and B. Denton. 2008. Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions* 40(9) 800–819.

Haas-Wilson, D. 2001. Arrow and the information market failure in health care: The changing content and sources of health care information. *Journal of Health Politics, Policy and Law* **26**(5) 1031–1044.

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

Integrated Healthcare Association. Pay-for-Performance Manuals and Operations: MY 2011 Measure Set. http://www.iha.org/pdfs_documents/p4p_california/ApprovedMY2011MeasureSet111610.pdf (accessed on March 15, 2011).

Institute of Medicine. 2001. *Crossing the Quality Chasm: A New Health System for the 21st Century.* Washington, D.C.: National Academy Press.

Kaarboe, O., and L. Siciliani. 2011. Multi-tasking, quality and pay for performance. *Health Economics* **20**(2) 225–238.

Kaiser Family Foundation. 2009. *Medicare Advantage Fact Sheet.* Washington, DC: Henry J. Kaiser Family Foundation.

Kim, S.-H., M.A. Cohen, and S. Netessine. 2007. Performance contracting in after-sales service supply chains. *Management Science* **53**(12) 1843–1858.

Leape L.L., and D.M. Berwick. 2005. Five years after *To Err Is Human*: What have we learned? *Journal of The American Medical Association* **293**(19) 2384–2390.

Lee, D.K.K., and S.A. Zenios. 2007. Evidence-based incentive systems with an application in health care delivery. Working Paper, Stanford University.

Lu M., and C. Donaldson. 2000. Performance-Based Contracts and Provider Efficiency: The State of the Art. *Disease Management & Health Outcomes* **7**(3) 127–137.

McGlynn, E.A., S.M. Asch, J. Adams, J. Keesey, J. Hicks, A. DeCristofaro, and E.A. Kerr. 2003. The quality of health care delivered to adults in the United States. *New England Journal of Medicine* **348**(26) 2635–2645.

Miraldo, M., L. Siciliani, and A. Street. 2011. Price adjustment in the hospital sector. *Journal of Health Economics* **30**(1) 112–125.

Mullen, K.J., R.G. Frank, and M.B. Rosenthal. 2010. Can you get what you pay for? Pay-for-performance and the quality of health care providers. *RAND Journal of Economics* **41**(1) 64–91.

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

Roland M. 2004. Linking physicians pay to the quality of care - A major experiment in the United Kingdon. *New England Journal of Medicine* **351**(14) 1448-1454.

Rosenthal M.B., R. Fernandopulle, H.R. Song, and B. Landon. 2004. Paying for quality: Providers incentives for quality improvement. *Health Affairs* **23**(2) 127-141.

Siciliani, L. 2007. Optimal contracts for health services in the presence of waiting times and asymmetric information. *The B.E. Journal of Economic Analysis & Policy* **7**(1) Article 40.

So, K.C., and 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.

Su, X., and S. A. Zenios. 2006. Recipient choice can address the efficiency-equity trade-off in kidney transplantation: A mechanism design problem. *Management Science* **52**(11) 1647–1660.

Tian, N., and Z.G. Zhang. 2006. *Vacation Queueing Models: Theory and Applications*. Springer Science+Business Media, LLC, New York.

Topkis, D.M. 1978. Minimizing a Submodular Function on a Lattice. *Operations Research* **26**(2) 305-321.

UK National Health Service. 2008. The standard NHS contracts for acute hospital, mental health, community and ambulance services and supporting guidance. http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_091451 (accessed on March 15, 2011).

Appendix A: Notation

Table 1 **Notation**

C	=	total number of daily service slots
A	=	number of daily service slots allocated for advance appointments
Z	=	total number of service slots made available for advance appointments on CaB
D_0	=	same-day demand
λ_0	=	expected value of same-day demand
λ	=	expected value of daily advance appointment requests
ρ	=	λ/A, average number of advance appointment requests per service slot
θ	=	fraction of dedicated patients
X	=	number of patients in the system
$\mathcal{L}$	=	length of patient waiting list
L_q	=	average length of patient waiting list
W_q	=	average patient waiting time
M	=	target waiting time (measured in days)
r	=	reimbursement per patient served
l	=	daily penalty cost incurred by the service provider for each patient on the waiting list
b	=	per-patient diverting cost
o	=	per-patient overtime cost
T	=	transfer payment between the purchaser and the provider
F	=	fixed payment under a threshold-penalty contract
K	=	penalty under a threshold-penalty contract
Π_a	=	expected profit for the provider
Π_p	=	expected cost for the purchaser
H	-	a superscript representing a provider with high overtime cost
L	-	a superscript representing a provider with low overtime cost

Appendix B: Proofs

Proof of Proposition 1

Some preliminary results on stochastic dominance for the modified $M/D/1$ queue are presented before we prove the results in Proposition 1. Let $X(t)$ be the number of patients in the appointment queue at time t and $S(t)$ the residual service time for a patient in service (if there is any). Then the system can be fully characterized by the two-dimensional state variable $(X(t), S(t))$. At any point in time, the length of the queue is $L_q(t) = (X(t) - 1)^+$ and the length of time an incoming patient will wait is $W_q(t) = (X(t) - 1)^+ + S(t)$.

We approximate the continuous-time system with a discrete-time system with time intervals of equal length $\delta = 1/N$, where N is a large positive integer. Then, each service time slot contains N successive intervals. The system state can be represented by (x, s), where x is the number of patients in the system and s is the number of residual service time intervals of the patient in service, $s = 0, 1, \ldots, N$. Note that we assume that $s = 0$ whenever $x = 0$. The Poisson arrival process is approximated by a Bernoulli process such that there is at most one patient arriving in each time interval. We assume that the service starts at the beginning of a time interval and occupies the entire time interval. If there is a patient on the waiting list at the beginning of service, his/her service starts immediately after the completion of the previous service. For instance, if the system is empty at the beginning of an interval, then the service of the first patient to arrive during that interval starts at the beginning of the following interval, at which point the system state changes from $(0,0)$ to $(1,N)$. Assuming that the system state at the beginning of a time interval is $(x,1)$, $x > 0$, and there is no patient arrival during this this time interval, the system state changes to $(x-1,N)$ at the end of this time interval. If, however, there is a patient arrival, the state changes to (x,N). Note that the state $(x,0)$ is observed only if $x = 0$. Similarly, if the system state at the beginning of a time interval is $(x,i), x > 0, i > 1$, the system state changes to $(x,i-1)$ if there is no patient arrival during that time interval, and to $(x,i-1)$ otherwise.

The two-dimensional state variable can be aggregated into a single state variable $i = N(x-1)^+ + s$, which represents the total number of time intervals an incoming patient should wait to be served. That is, given any state i, the number of patients in the system (including the one in service) is $\lceil i/N \rceil$ (the nearest integer greater than or equal to i/N) and the number of time intervals of the residual service time of the patient in service is $i - N\lfloor i/N \rfloor$ where $\lfloor i/N \rfloor$ represents the nearest integer less than or equal to i/N. Note that $i > 0$ implies that $x > 0$, and vice versa. Let $\rho = \lambda/A$. Given the policy parameter, Z, if the number of patients in the system, $\lceil i/N\ rceil$, is less than Z, then with a probability of $\rho\delta$ an appointment request will arrive and a patient will join the waiting list, and with a probability of $1-\rho\delta$ there is no patient arrival. If $\lceil i/N \rceil \geq Z$, then with probability $\theta\rho\delta$ a patient will join the waiting list, and with probability $1-\theta\rho\delta$ there is no patient arriving in this time interval. We assume that service for each patient starts only in the beginning of an interval. The transition matrix for the time-discretized Markov chain is represented by $\Pi_N = (\pi_{i,j}(\delta))$, where

$$\pi_{i,j}(\delta) = \begin{cases} 1-\rho\delta & \text{if } i/N \leq Z-1, j=(i-1)^+, \\ \rho\delta & \text{if } i/N \leq Z-1, j=(i-1)^+ + N, \\ 1-\theta\rho\delta & \text{if } i/N > Z-1, j=(i-1)^+, \\ \theta\rho\delta & \text{if } i/N > Z-1, j=(i-1)^+ + N, \\ 0 & \text{otherwise} \end{cases}$$

for $i, j \in \mathbb{N}$ or

$$\Pi_N = \begin{pmatrix} 1-\rho\delta & 0 & 0 & \cdots & \rho\delta & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots \\ 1-\rho\delta & 0 & 0 & \cdots & \rho\delta & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots \\ 0 & 1-\rho\delta & 0 & \cdots & 0 & \rho\delta & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots \\ 0 & 0 & 1-\rho\delta & \cdots & 0 & 0 & \rho\delta & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1-\rho\delta & 0 & 0 & \cdots & \rho\delta & 0 & 0 & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 1-\theta\rho\delta & 0 & \cdots & 0 & \theta\rho\delta & 0 & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \iddots & 0 & 0 & 1-\theta\rho\delta & \cdots & 0 & 0 & \theta\rho\delta & \cdots \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \quad \text{(B1)}$$

Let I represent the aggregated system state at any random time. Note that we require $\rho\theta < 1$, which ensures the existence of the stationary distribution of I. Denote the stationary distribution by $Q = [q_0, q_1, \cdots]$, where q_i is the stationary probability for the system state i. Note that $Q = Q\Pi_N = \lim_{n\to\infty} Q\Pi_N^n$.

A sequence $X = [x_0, x_1, \cdots]'$ is called an increasing sequence if $x_i \leq x_{i+1}$ for any $i = 0, 1, \ldots$. The next lemma shows that matrix Π_N maps an increasing sequence into an increasing sequence.

Lemma B1. *For any increasing sequence* $X = [x_0, x_1, \cdots]'$, $Y = \Pi_N X$ *is also an increasing sequence.*

Proof. It is straightforward to see that if $(i+1)/N \leq Z-1$ or $(i-1)/N \geq Z-1$, then

$$\pi_{i,(i-1)^+} x_{(i-1)^+} + \pi_{i,(i-1)^+ + N} x_{(i-1)^+ + N} \leq \pi_{i+1,i} x_i + \pi_{i+1,i+N} x_{i+N}, \quad \text{(B2)}$$

where $\pi_{i,(i-1)^+ + N} = \pi_{i+1,i+N}$ and is equal to $\rho\delta$ or $\theta\rho\delta$, and $\pi_{i,(i-1)^+} = \pi_{i+1,i}$ and is equal to $1-\rho\delta$ or $1-\theta\rho\delta$. Otherwise, $i/N = Z-1$, then

$$[\pi_{i,i-1} x_{i-1} + \pi_{i,i-1+N} x_{i-1+N}] - [\pi_{i+1,i} x_i + \pi_{i+1,i+N} x_{i+N}] \quad \text{(B3)}$$

$$
\begin{aligned}
&= [(1-\rho\delta)x_{i-1} + \rho\delta x_{i-1+N}] - [(1-\theta\rho\delta)x_i + \theta\rho\delta x_{i+N}] && \text{(B4)}\\
&= [(1-\rho\delta)x_{i-1} + \rho\delta x_{i-1+N}] - [(1-\rho\delta)x_i + \rho\delta x_{i+N} + (1-\theta)\rho\delta(x_i - x_{i+N})] && \text{(B5)}\\
&\le 0, && \text{(B6)}
\end{aligned}
$$

where the last inequality follows from the fact that X is an increasing sequence. Then, $\Pi_N X$ is also an increasing sequence. □

Consider two parameter triples, (θ_k, ρ_k, Z_k), $k = 1, 2$. Let $X^k = [x_1^k, x_2^k, \cdots]$, $k = 1, 2$, be two increasing sequences such that $X^1 \le X^2$, i.e., $x_i^1 \le x_i^2$ for all i. Let Π_N^k be the respective transition matrices. The next lemma shows the monotone preservation property of Π_N.

Lemma B2. *If $\theta_1 \le \theta_2$, $\rho_1 \le \rho_2$ and $Z_1 \le Z_2$, then $\Pi_N^1 X^1 \le \Pi_N^2 X^2$.*

Proof. We have

$$
\begin{aligned}
\Pi^1 X^1 &= [\pi^1_{0,0}x^1_0 + \pi^1_{0,N}x^1_N, \pi^1_{1,0}x^1_0 + \pi^1_{1,N}x^1_N, \cdots, \pi^1_{N(Z_1-1)+1,N(Z_1-1)}x^1_{N(Z_1-1)} + \pi^1_{NZ_1+1,NZ_1}x^1_{NZ_1},\\
&\quad \cdots, \pi^1_{N(Z_2-1)+1,N(Z_2-1)}x^1_{N(Z_2-1)} + \pi^1_{NZ_2+1,NZ_2}x^1_{NZ_2}, \cdots]'\\
&\le [\pi^1_{0,0}x^2_0 + \pi^1_{0,N}x^2_N, \pi^1_{1,0}x^2_0 + \pi^1_{1,N}x^2_N, \cdots, \pi^1_{N(Z_1-1)+1,N(Z_1-1)}x^2_{N(Z_1-1)} + \pi^1_{NZ_1+1,NZ_1}x^2_{NZ_1},\\
&\quad \cdots, \pi^1_{N(Z_2-1)+1,N(Z_2-1)}x^2_{N(Z_2-1)} + \pi^1_{NZ_2+1,NZ_2}x^2_{NZ_2}, \cdots]'\\
&\le [\pi^2_{0,0}x^2_0 + \pi^2_{0,N}x^2_N, \pi^2_{1,0}x^2_0 + \pi^2_{1,N}x^2_N, \cdots, \pi^2_{N(Z_1-1)+1,N(Z_1-1)}x^2_{N(Z_1-1)} + \pi^2_{NZ_1+1,NZ_1}x^2_{NZ_1},\\
&\quad \cdots, \pi^2_{N(Z_2-1)+1,N(Z_2-1)}x^2_{N(Z_2-1)} + \pi^2_{NZ_2+1,NZ_2}x^2_{NZ_2}, \cdots]'\\
&= \Pi^2 X^2,
\end{aligned}
$$

where the first inequality follows from the assumption that $X^1 \le X^2$ (i.e., $x_i^1 \le x_i^2$ for all i) and the second from the fact that $\pi^1_{i,(i-1)^+} \ge \pi^2_{i,(i-1)^+}$ and $\pi^1_{i,(i-1)^++N} \le \pi^2_{i,(i-1)^++N}$, and therefore

$$
\begin{aligned}
\pi^1_{i,(i-1)^+}x^2_{(i-1)^+} + \pi^1_{i,(i-1)^++N}x^2_{(i-1)^++N} &= x^2_{(i-1)^+} + \pi^1_{i,(i-1)^++N}(x^2_{(i-1)^++N} - x^2_{(i-1)^+})\\
&\le x^2_{(i-1)^+} + \pi^2_{i,(i-1)^++N}(x^2_{(i-1)^++N} - x^2_{(i-1)^+})\\
&= \pi^2_{i,(i-1)^+}x^2_{(i-1)^+} + \pi^2_{i,(i-1)^++N}x^2_{(i-1)^++N}.
\end{aligned}
$$

□

Let I^k be the number of time intervals an incoming patient will wait to be served corresponding to $(\theta_k, \rho_k, Z_k), k = 1, 2$. We say that I^1 is stochastically smaller (denoted by $\le_{st}$) than I^2 if for any increasing function h, $E[h(I^1)] \le E[h(I^2)]$. Let X_N^k be the number of patients in the system corresponding to (θ_k, ρ_k, Z_k). The next lemma proves that the stationary distributions of I and X_N are stochastically monotone in θ, ρ, and Z.

Lemma B3. *If $\theta_1 \le \theta_2$, $\rho_1 \le \rho_2$ and $Z_1 \le Z_2$, then $I^1 \le_{st} I^2$ and $X_N^1 \le_{st} X_N^2$. In addition, $W_q^1 \le W_q^2$, $L_q^1 \le L_q^2$ and $Pr(X_N^1 \ge Z_1) \ge Pr(X_N^2 \ge Z_2)$.*

Proof. Let $h(\cdot):\mathbb{Z}_+ \to \mathbb{R}$ be any increasing function. Then $Y = [h(0), h(1), \cdots]'$ is an increasing sequence. Applying Lemma B2 yields $\Pi^1_N Y \leq \Pi^2_N Y$. For any integer $n \geq 1$, we have $(\Pi^1_N)^n \cdot Y \leq (\Pi^2_N)^n \cdot Y$.

Let Q^k be the stationary distribution of the number of patients in the system corresponding to (θ_k, ρ_k, Z_k). Then,

$$Q^k = Q^k \Pi^k_N = \lim_{n\to\infty} Q^0 (\Pi^k_N)^n,$$

where Q^0 can be any starting distribution. Then, we have

$$Q^1 \cdot Y = \lim_{n\to\infty} Q^0 (\Pi^1_N)^n \cdot Y \leq \lim_{n\to\infty} Q^0 (\Pi^2_N)^n \cdot Y = Q^2 \cdot Y.$$

That is, $E[h(I^1)] \leq E[h(I^2)]$ where $I^k, k = 1,2$ represent aggregated system states. Then, $I^1 \leq_{st} I^2$. Note that the number of patients in the system, $X^k_N = \lceil I/N \rceil$, is an increasing function of I^k. Then, $X^1_N \leq_{st} X^2_N$. As $W^k_q = E[I^k \delta]$, the stochastic monotonicity implies that $W^1_q \leq W^2_q$. Similar, as $L^k_q = E[(X^k_N - 1)^+]$, then $L^1_q \leq L^2_q$. The stochastic monotonicity also implies that $Pr(X^1_N > 0) \leq Pr(X^2_N > 0)$. Note that under equilibrium the average arrival rate is $\rho(1-(1-\theta)Pr(X^k_N \geq Z)) = \rho(1-(1-\theta)Pr(I^k \geq NZ))$. By the conservation law, $Pr(X^k_N > 0) = \rho(1-(1-\theta)Pr(X^k_N \geq Z^k))$. Then, for any $\theta < 1$, $Pr(X^1_N > 0) \leq Pr(X^2_N > 0)$ implies that $Pr(X^1_N \geq Z^1) \geq Pr(X^2_N \geq Z^2)$. □

As $N \to \infty$ ($\delta \to 0$), the Bernoulli process converges to the Poisson process and the discrete-time system converges to the continuous-time system. Then the stationary distribution of X_N converges to the stationary distribution of $X(t)$, which, consequently, has the stochastically monotone properties characterized in B3. Then, the monotone properties in Proposition 1 follow. □

Proof of Lemma 1

Proposition 1 shows that $W_q(A,Z)/A$ is monotone increasing in Z and monotone decreasing in A. Then for any $\theta \in [0,1]$, any $Z \geq 0$, and any $A \geq A^*$, we have

$$\begin{aligned}\frac{W_q(A,Z)}{A} &\leq \frac{W_q(A,\infty)}{A}\\ &\leq \frac{W_q(A^*,\infty)}{A^*}\\ &= \frac{\lambda}{2A^*(A^*-\lambda)}\\ &= M,\end{aligned} \tag{B7}$$

where the first equality follows from (6) and the second from the definition of A^*. This shows that the service level constraint is satisfied. □

Proof of Proposition 2

(a) Proposition 1 states that $Pr\left(X(A^t, Z^t) \geq Z^t\right)$ is decreasing in Z^t for any given A^t, and decreasing in A^t for any given Z^t. Then, the objective function of the first-best problem (21) is decreasing in Z^t, which implies that the service level constraint must be satisfied as tightly as possible at the optimal solution, i.e., $W_q\left(A^t_{\mathrm{FB}}, Z^t_{\mathrm{FB}}\right)/A^t_{\mathrm{FB}} \leq M$ while $W_q\left(A^t_{\mathrm{FB}}, Z^t_{\mathrm{FB}}+1\right)/A^t_{\mathrm{FB}} > M$. Now, consider the first-best solution $(T^t_{\mathrm{FB}}, Z^t_{\mathrm{FB}}, A^t_{\mathrm{FB}})$. Suppose that $\Pi^t_a\left(T^t_{\mathrm{FB}}, A^t_{\mathrm{FB}}, Z^t_{\mathrm{FB}}\right) > 0$. Then, since both T^t and Π^t_a are monotone increasing in T^t, we can improve the objective function by lowering T^t_{FB} without violating the individual rationality constraint. Thus, $\Pi^t_a\left(T^t_{\mathrm{FB}}, A^t_{\mathrm{FB}}, Z^t_{\mathrm{FB}}\right)$ has to be equal to 0. Then, (21) is obtained by replacing T^t in the objective function of the purchaser's problem by its expression from (10).

(b) As Proposition 1 states, $W_q\left(A^t, Z^t\right)/A^t$ is increasing in Z^t and decreasing in A^t. This, in turn, implies that $Z^t_M(A^t)$ is increasing in A^t. Let A^t_{FB} be the solution to (21). Observe that the objective function in this problem is supermodular in (o^t, A^t). Then, applying Theorem 6.3 from Topkis (1978), we obtain that A^t_{FB} is non-increasing in o^t, which implies that $A^H_{FB} \leq A^L_{FB}$. Moreover, $Z^H_{FB} = Z^t(A^H_{FB}) \leq Z^t(A^L_{FB}) = Z^L_{FB}$. Similarly, the objective function is submodular in (b, A^t), and, consequently, A^t_{FB} and Z^t_{FB} are both non-decreasing in b.

(c) The result follows from the definition of the performance-based contract and the formula for T^t_{FB}. □

Proof of Corollary 1

(a) For $o^t = 0$, (21) reflects the minimization of $Pr\left(X\left(A^t, Z^t\right) \geq Z^t\right)$. For any finite Z^t, as follows from Proposition 1, this objective is minimized by setting $A^t = C$. Note that for $Z^t \to +\infty$, the appointment dynamics is identical to one of the $M/D/1$ queue, and $\lim_{Z^t \to +\infty} Pr\left(X\left(C, Z^t\right) \geq Z^t\right) = 0$, as long as the corresponding $M/D/1$ system is stable, i.e., as long as $C > \lambda$. This last condition is implied by (6), which also ensures that the waiting-time requirement is satisfied.

(b) For $b = 0$, the objective function in (21), for given Z^t, is minimized by setting A^t to the smallest possible value compatible with the service level constraint $W(A^t, Z^t)/A^t$. Since Proposition 1 shows that $W(A^t, Z^t)/A^t$ is an increasing function of Z^t and a decreasing function of A^t, the value of Z^t has to be set at the lowest possible value. For $Z^t = 0$, the appointment dynamics becomes that of an $M/D/1$ queue with a Poisson arrival rate of $\theta\lambda$, and the patient waiting time constraint becomes

$$\frac{\theta\lambda}{2A^t(A^t - \theta\lambda)} \leq M, \tag{B8}$$

which is equivalent to

$$A^t \geq \frac{\theta\lambda}{2} + \sqrt{\frac{\theta^2\lambda^2}{4} + \frac{\theta\lambda}{2M}}. \tag{B9}$$

Thus, $A^t_{\text{FB}} = \frac{\theta\lambda}{2} + \sqrt{\frac{\theta^2\lambda^2}{4} + \frac{\theta\lambda}{2M}}$.

(c) For $\theta = 1$, the optimization objective is the same as in part (b). At the same time, $\theta = 1$ also implies that the appointment dynamics becomes that of an $M/D/1$ queue with a Poisson arrival rate of λ, irrespective of the chosen value of Z^t. Using the same arguments, we obtain $A^t_{\text{FB}} = \frac{\lambda}{2} + \sqrt{\frac{\lambda^2}{4} + \frac{\lambda}{2M}}$. □

Proof of Lemma 2

Note that the expected transfer payment incurred under the infinite-horizon (IH) policy can be expressed as $T_{\text{IH}} = (po^H + (1-p)o^L)E_{D_0}[(D_0 - C + A^*)^+]$. In order to prove the statement of the Lemma, we need to establish the lower bound on the optimal values of the optimization problem (20)-(21), T^t_{FB}. Such lower bound can be obtained by dropping the constraint on Z^t in (20). The solution to the resulting optimization problem is $A^t = \frac{\theta\lambda}{2} + \sqrt{\frac{\theta^2\lambda^2}{4} + \frac{\theta\lambda}{2M}}$ and $Z^T = \infty$, with the corresponding optimal objective function value $o^t E_{D_0}\left[\left(D_0 - C + \frac{\theta\lambda}{2} + \sqrt{\frac{\theta^2\lambda^2}{4} + \frac{\theta\lambda}{2M}}\right)^+\right]$. Thus, $T^t_{\text{FB}} \geq o^t E_{D_0}\left[\left(D_0 - C + \frac{\theta\lambda}{2} + \sqrt{\frac{\theta^2\lambda^2}{4} + \frac{\theta\lambda}{2M}}\right)^+\right]$. Using $T_{\text{IH}} = (po^H + (1-p)o^L)E_{D_0}\left[\left(D_0 - C + \frac{\lambda}{2} + \sqrt{\frac{\lambda^2}{4} + \frac{\lambda}{2M}}\right)^+\right]$, we obtain the result. □

Proof of Proposition 3

First, note that the provider's objective function $\Pi^t_a(r^t, l^t, A)$ is concave in A. Indeed,

$$\frac{\partial \Pi^t_a}{\partial A} = l^t \frac{\lambda}{2}\left[\frac{1}{(A-\lambda)^2} - \frac{1}{A^2}\right] - o^t(1 - F_{D_0}(C - A)), \tag{B10}$$

and

$$\frac{\partial^2 \Pi^t_a}{\partial A^2} = l^t \lambda \left[\frac{1}{A^3} - \frac{1}{(A-\lambda)^3}\right] - o^t f_{D_0}(C - A) \leq 0. \tag{B11}$$

The first-order optimality condition for the provider of type t is:

$$-l^t \frac{\lambda}{2}\left[-\frac{1}{(A-\lambda)^2} + \frac{1}{A^2}\right] - o^t(1 - F_{D_0}(C - A)) = l^t \frac{\lambda^2(2A - \lambda)}{2A^2(A-\lambda)^2} - o^t(1 - F_{D_0}(C - A)) = 0. \tag{B12}$$

Under the contract (37)-(38), A^* defined in (8) satisfies the above first-order optimality condition as well as simple bounds constraints. The concavity of the objective function implies that A^* is an optimal solution for the provider's problem.

Next, it is easy to check that given contract (37)-(38), the optimal solution A^* for the provider satisfies the service level constraint and gives an objective function value $o^t E_{D_0}[(D_0 - C + A^*)^+]$, which is equal to the optimal objective function value for the purchaser in the first-best solution. Hence, we have proved that the contract (37)-(38) achieves the first-best outcome. □

Proof of Proposition 4

(a) First, note that the purchaser's optimization problem can be reformulated as

$$\begin{aligned}
\min_{T^t,A^t,Z^t,t\in\{H,L\}} \quad & \left(pT^H+(1-p)T^L\right), & \text{(B13)}\\
\text{s.t.} \quad & (A^t,Z^t)\in\mathcal{R}\left(M,C,\theta,\lambda\right), t\in\{H,L\}, & \text{(B14)}\\
& \Pi_a^{tt}\left(T^t,A^t,Z^t\right)\geq 0, t\in\{H,L\}, & \text{(B15)}\\
& \Pi_a^{tt}\left(T^t,A^t,Z^t\right)\geq \Pi_a^{ts}\left(T^s,A^s,Z^s\right), t,s\in\{H,L\}, s\neq t, & \text{(B16)}
\end{aligned}$$

where

$$\begin{aligned}
\Pi_a^{tt}\left(T^t,A^t,Z^t\right) &= T^t - o^t E_{D_0}[(D_0-C+A^t)^+] - b\lambda(1-\theta)Pr\left((X(A^t,Z^t)\geq Z^t\right), t\in\{H,L\},\\
\Pi_a^{ts}\left(T^s,A^s,Z^s\right) &= T^s - o^t E_{D_0}[(D_0-C+A^s)^+] - b\lambda(1-\theta)Pr\left((X(A^s,Z^s)\geq Z^s\right),\\
& \quad t,s\in\{H,L\}, s\neq t. \qquad \text{(B17)}
\end{aligned}$$

Note that the constraint (B16) for $t=L$ and $s=H$,

$$\begin{aligned}
& T^L - o^L E_{D_0}[(D_0-C+A^L)^+] - b\lambda(1-\theta)Pr\left((X(A^L,Z^L)\geq Z^L\right)\geq\\
& T^H - o^L E_{D_0}[(D_0-C+A^H)^+] - b\lambda(1-\theta)Pr\left((X(A^H,Z^H)\geq Z^H\right), \qquad \text{(B18)}
\end{aligned}$$

is equivalent to

$$\Pi_a^{LL}\left(T^L,A^L,Z^L\right)\geq \Pi_a^{HH}\left(T^H,A^H,Z^H\right)+(o^H-o^L)E_{D_0}[(D_0-C+A^H)^+]. \qquad \text{(B19)}$$

Then, since $\Pi_a^{HH}\left(T^H,A^H,Z^H\right)\geq 0$, (B19) implies that $\Pi^{LL}\left(T^L,A^L,Z^L\right)>0$. Note that $\Pi_a^{HH}\left(T^H,A^H,Z^H\right)=0$ at the optimum: otherwise, the purchaser can reduce the objective function value by decreasing both T^H and T^L by the same amount without violating (B15) and (B16). Thus, the optimal transfer payment designed for the high-cost provider is equal to the sum of the overtime and the patient-diverting costs:

$$T^H = o^H E_{D_0}[(D_0-C+A^H)^+] + b\lambda(1-\theta)Pr\left(X(A^H,Z^H)\geq Z^H\right). \qquad \text{(B20)}$$

The inequality (B18) (or equivalently (B19)) must be binding at the optimum, i.e., the expected net payoff of the low-cost provider is equal to the information rent. Otherwise, the purchaser can reduce T^L, without violating other constraints, until (B18) or (B19) is binding. Thus, at the optimum,

$$\begin{aligned}
T^L &= o^L E_{D_0}[(D_0-C+A^L)^+] + b\lambda(1-\theta)Pr\left((X(A^L,Z^L)\geq Z^L\right)\\
&\quad + (o^H-o^L)E_{D_0}[(D_0-C+A^H)^+]. \qquad \text{(B21)}
\end{aligned}$$

As $\Pi_a^{HH}(T^H,A^H,Z^H)=0$ and $\Pi_a^{LL}(T^L,A^L,Z^L)=(o^H-o^L)E_{D_0}[(D_0-C+A^H)^+]$, the constraint

$$\Pi_a^{HH}\left(T^H,A^H,Z^H\right)\geq\Pi_a^{LL}\left(T^L,A^L,Z^L\right)+(o^L-o^H)E_{D_0}[(D_0-C+A^L)^+] \tag{B22}$$

is equivalent to

$$(o^H-o^L)(E_{D_0}[(D_0-C+A^H)^+]-E_{D_0}[(D_0-C+A^L)^+])\leq 0, \tag{B23}$$

which implies that $A^H\leq A^L$. Finally, (46) is obtained by replacing T^H and T^L in the objective function of the purchaser's problem by their expressions from (B20) and (B21).

(b) The results follow from (B20) and (B21).

(c) Since $\hat{o}^H>o^H>o^L$, the result of part (b) of Proposition 2 implies that the optimal solution to (46) automatically satisfies $A_{\mathrm{SB}}^H\leq A_{\mathrm{FB}}^H\leq A_{\mathrm{SB}}^L=A_{\mathrm{FB}}^L$. Given that the service-level constraints are binding at both first-best and second-best solutions and W_q/A is increasing Z and decreasing in A, we have $Z_{\mathrm{SB}}^H\leq Z_{\mathrm{FB}}^H\leq Z_{\mathrm{SB}}^L=Z_{\mathrm{FB}}^L$.

(d) The result follows from the definitions for the payments T^t when T^t takes the linear performance-based contract and part (b) of this proposition. □

Proof of Proposition 5

(a), (b) Note that $\frac{\partial\Pi_a^{ts}(r^s,l^s,A^{ts})}{\partial o^t}=-E_{D_0}[(D_0-C+A^{ts})^+]$ is decreasing in A^{ts}. Thus, $\Pi_a^{ts}(r^s,l^s,A^{ts})$ is submodular in (o^t,A^{ts}). Then, as follows from Theorem 6.3 in Topkis (1978), the maximizer of Π_a^{ts} in terms of A^{ts} is increasing in o^t, which implies that $A_{\mathrm{PA}}^{LH}\geq A_{\mathrm{PA}}^{HH}$ and $A_{\mathrm{PA}}^{LL}\geq A_{\mathrm{PA}}^{HL}$.

If $\lambda<A_{\mathrm{PA}}^{HH}<C$, then A_{PA}^{HH} is an interior optimum of provider's objective function, and

$$\frac{\lambda l^H(2A_{\mathrm{PA}}^{HH}-\lambda)}{2(A_{\mathrm{PA}}^{HH}(A_{\mathrm{PA}}^{HH}-\lambda))^2}-o^H(1-F_{D_0}(C-A_{\mathrm{PA}}^{HH}))=0. \tag{B24}$$

Then,

$$\frac{\lambda l^H(2A_{\mathrm{PA}}^{HH}-\lambda)}{2(A_{\mathrm{PA}}^{HH}(A_{\mathrm{PA}}^{HH}-\lambda))^2}-o^L(1-F_{D_0}(C-A_{\mathrm{PA}}^{HH}))>0. \tag{B25}$$

The concavity of $\Pi_a^{ts}(r^s,l^s,A^{ts})$ with respect to A^{ts} implies that

$$\frac{\partial\Pi_a^{LH}}{\partial A}=\frac{\lambda l^H(2A-\lambda)}{2(A(A-\lambda))^2}-o^L(1-F_{D_0}(C-A)) \tag{B26}$$

is decreasing in A, which, in turn, implies that $A_{\mathrm{PA}}^{LH}>A_{\mathrm{PA}}^{HH}$. Similarly, if $\lambda<A_{\mathrm{PA}}^{HL}<C$, then $A_{\mathrm{PA}}^{LL}>A_{\mathrm{PA}}^{HL}$.

(c) As $\lambda/(2A^{ts}(A^{ts}-\lambda))$ is decreasing in A^{ts}, the provider's profit $\Pi_a^{ts}(r^s,l^s,A^{ts})$ is supermodular in (l^s,A^{ts}). Then, as follows from Theorem 6.3 in Topkis (1978), A^{ts} is increasing in l^s. Proposition 3 states that A^* is the optimal solution for the provider's optimization problem for $l^s=\tilde{l}^t$. Thus, $A_{\mathrm{PA}}^{ts}\leq A^*$ if and only if $l^s\leq\tilde{l}^t$. In particular, from $A^*<C$ and $\Pi_a^{ts}(r^s,l^s,A^{ts})$ being strictly concave in A^{ts}, it follows that $A^{ts}>A^*$ for $l^s>\tilde{l}^t$. □

Proof of Proposition 6

(a) As follows from Proposition 5, $l^H \geq \tilde{l}^H$ and $l^L \geq \tilde{l}^L$, since, otherwise, the agent will choose A_{PA}^{HH} or A_{PA}^{LL} that are less than A^*, and the patient waiting-time constraint would be violated.

(b) Note that

$$\begin{aligned}\Pi_a^{LH}(r^H, l^H, A_{\mathrm{PA}}^{LH}) &= r^H(\lambda+\lambda_0) - l^H\lambda/(2A_{\mathrm{PA}}^{LH}(A_{\mathrm{PA}}^{LH}-\lambda)) - o^L E_{D_0}[(D_0 - C + A_{\mathrm{PA}}^{LH})^+] \\ &\geq r^H(\lambda+\lambda_0) - l^H\lambda/(2A_{\mathrm{PA}}^{HH}(A_{\mathrm{PA}}^{HH}-\lambda)) - o^L E_{D_0}[(D_0 - C + A_{\mathrm{PA}}^{HH})^+] \\ &= \Pi_a^{HH}(r^H, l^H, A_{\mathrm{PA}}^{HH}) + (o^H - o^L)E_{D_0}[(D_0 - C + A_{\mathrm{PA}}^{HH})^+] \\ &> \Pi_a^{HH}(r^H, l^H, A_{\mathrm{PA}}^{HH}),\end{aligned} \tag{B27}$$

where the first inequality follows from the optimality of A_{PA}^{LH} and the second one from the condition $o^L < o^H$. Using (61) and (58), we get

$$\Pi_a^{LL}(r^L, l^L, A_{\mathrm{PA}}^{LL}) \geq \Pi_a^{LH}(r^H, l^H, A_{\mathrm{PA}}^{LH}) > \Pi_a^{HH}(r^H, l^H, A_{\mathrm{PA}}^{HH}) \geq 0, \tag{B28}$$

which implies that $\Pi_a^{LL}(r^L, l^L, A_{\mathrm{PA}}^{LL}) > 0$. Thus, (59) is not binding at the optimum and can be ignored.

Further, observe that the terms $T^{LL}(r^L, l^L, A_{\mathrm{PA}}^{LL})$ and $T^{HH}(r^H, l^H, A_{\mathrm{PA}}^{HH})$ from the objective function of the principal are strictly increasing in r^L and r^H, respectively. Then, (58) will have to be binding at the optimum: otherwise, the principal can reduce r^H and r^L by the same amount until (58) is binding while leaving the constraints (60) and (61) unaffected. Similarly, (61) must be binding at the optimum: otherwise, the principal can reduce the r^L until (61) is binding without impacting (58).

Now we can show that $A_{\mathrm{PA}}^{LL} > A^*$. Indeed, suppose that $A_{\mathrm{PA}}^{LL} = A^*$. Then, as follows from Proposition 5, $A_{\mathrm{PA}}^{HL} < A_{\mathrm{PA}}^{LL}$. Since (58) is binding at the optimum, from (60) we have

$$\begin{aligned}0 &\geq r^L(\lambda+\lambda_0) - l^L\lambda/(2A_{\mathrm{PA}}^{HL}(A_{\mathrm{PA}}^{HL}-\lambda)) - o^H E_{D_0}[(D_0 - C + A_{\mathrm{PA}}^{HL})^+] \\ &> r^L(\lambda+\lambda_0) - l^L\lambda/(2A_{\mathrm{PA}}^{LL}(A_{\mathrm{PA}}^{LL}-\lambda)) - o^H E_{D_0}[(D_0 - C + A_{\mathrm{PA}}^{LL})^+] \\ &= \Pi_a^{LL}(r^L, l^L, A_{\mathrm{PA}}^{LL}) - (o^H - o^L)E_{D_0}[(D_0 - C + A_{\mathrm{PA}}^{LL})^+],\end{aligned} \tag{B29}$$

where the first inequality follows from (60) and the fact that (58) is binding, and the second is due to the strict concavity of $\Pi_a^{HL}(r^L, l^L, A^{HL})$ with respect to A^{HL} and the optimality of A_{PA}^{HL}. The above inequalities imply that $\Pi_a^{LL}(r^L, l^L, A_{\mathrm{PA}}^{LL}) < (o^H - o^L)E_{D_0}[(D_0 - C + A_{\mathrm{PA}}^{LL})^+] = (o^H - o^L)E_{D_0}[(D_0 - C + A^*)^+]$. However, since (61) is binding and (B27) holds, we get

$$\Pi_a^{LL}(r^L, l^L, A_{\mathrm{PA}}^{LL}) = \Pi_a^{LH}(r^H, l^H, A_{\mathrm{PA}}^{LH})$$

$$
\begin{aligned}
&\geq \Pi_a^{HH}(r^H, l^H, A_{\mathrm{PA}}^{HH}) + (o^H - o^L)E_{D_0}[(D_0 - C + A_{\mathrm{PA}}^{HH})^+] \\
&= (o^H - o^L)E_{D_0}[(D_0 - C + A_{\mathrm{PA}}^{HH})^+] \\
&\geq (o^H - o^L)E_{D_0}[(D_0 - C + A^*)^+], \qquad \text{(B30)}
\end{aligned}
$$

where the equality is due to the fact that (58) is binding. We have a contradiction. Thus, $A_{\mathrm{PA}}^{LL} > A^*$.

(c) We next establish that $\Pi_a^{LL}(r^L, l^L, A_{\mathrm{PA}}^{LL}) > (o^H - o^L)E_{D_0}[(D_0 - C + A^*)^+]$. If $A_{\mathrm{PA}}^{HH} > A^*$, then

$$
\Pi_a^{LL}(r^L, l^L, A_{\mathrm{PA}}^{LL}) \geq (o^H - o^L)E_{D_0}[(D_0 - C + A_{\mathrm{PA}}^{HH})^+] > (o^H - o^L)E_{D_0}[(D_0 - C + A^*)^+]. \quad \text{(B31)}
$$

On the other hand, if $A_{\mathrm{PA}}^{HH} = A^*$, then by Proposition 5, $A_{\mathrm{PA}}^{LH} > A_{\mathrm{PA}}^{HH}$. Then,

$$
\begin{aligned}
\Pi_a^{LL}(r^L, l^L, A_{\mathrm{PA}}^{LL}) &= \Pi_a^{LH}(r^H, l^H, A_{\mathrm{PA}}^{LH}) \\
&> \Pi_a^{LH}(r^H, l^H, A_{\mathrm{PA}}^{HH}) \\
&= (o^H - o^L)E_{D_0}[(D_0 - C + A_{\mathrm{PA}}^{HH})^+] \\
&= (o^H - o^L)E_{D_0}[(D_0 - C + A^*)^+]. \qquad \text{(B32)}
\end{aligned}
$$

Here, the first equality holds because (61) is binding at the optimum, while the inequality holds because $\Pi_a^{LH}(r^H, l^H, A^{LH})$ is strictly concave in A^{LH} and A_{PA}^{LH} is the unique optimal point. The second equality holds due to the fact that (58) is binding at the optimum.

(d) Using (B32), we get

$$
\begin{aligned}
&pT^H(r^H, l^H, A_{\mathrm{PA}}^{HH}) + (1-p)T^L(r^L, l^L, A_{\mathrm{PA}}^{LL}) \\
&= p(\Pi_a^{HH}(r^H, l^H, A_{\mathrm{PA}}^{HH}) + o^H E_{D_0}[(D_0 - C + A_{\mathrm{PA}}^{HH})^+]) + (1-p)(\Pi_a^{LL}(r^L, l^L, A_{\mathrm{PA}}^{LL}) + o^L E_{D_0}[(D_0 - C + A_{\mathrm{PA}}^{LL})^+]) \\
&> p(o^H E_{D_0}[(D_0 - C + A^*)^+]) + (1-p)((o^H - o^L)E_{D_0}[(D_0 - C + A^*)^+] + o^L E_{D_0}[(D_0 - C + A^*)^+]) \\
&= o^H E_{D_0}[(D_0 - C + A^*)^+]. \qquad \text{(B33)}
\end{aligned}
$$

□

Proof of Proposition 7

If a provider of type t chooses A^t and Z^t that violate the service-level constraint, then the definition of K implies that the objective function value at (A^t, Z^t) for the provider is

$$
F^t - K - o^t E_{D_0}[(D_0 - C + A^t)^+] - b\lambda(1-\theta)Pr(X(A^t, Z^t) \geq Z^t), \quad \text{(B34)}
$$

which is smaller than

$$
o^t E_{D_0}[(D_0 - C + \theta\lambda)^+] - o^t E_{D_0}[(D_0 - C + A^t)^+] - b\lambda(1-\theta)Pr(X(A^t, Z^t) \geq Z^t). \quad \text{(B35)}
$$

The latter is a negative number because $A^t \geq \theta\lambda$, which is a necessary condition to ensure the stability of the queueing system. On the other hand, if a provider of type t chooses $A^t = A_{\mathrm{FB}}^t$ and

$Z^t = Z^t_{\mathrm{FB}}$, then the service-level constraint is not violated, the transfer payment is F^t, and the objective function value at $(A^t_{\mathrm{FB}}, Z^t_{\mathrm{FB}})$ for the provider is zero. Therefore, $(A^t_{\mathrm{FB}}, Z^t_{\mathrm{FB}})$ is a better solution than any solution (A^t, Z^t) that violates the service-level constraint for the provider of type t.

Then, the problem for the provider of type-t is equivalent to

$$\max_{(A^t,Z^t)\in\mathcal{R}(M,C,\theta,\lambda)} \left(F^t - o^t E_{D_0}[(D_0 - C + A^t)^+] - b\lambda(1-\theta)Pr(X(A^t,Z^t)\geq Z^t)\right). \tag{B36}$$

Since F^t is constant, the provider's problem is further equivalent to

$$\min_{(A^t,Z^t)\in\mathcal{R}(M,C,\theta,\lambda)} \left(o^t E_{D_0}[(D_0 - C + A^t)^+] + b\lambda(1-\theta)Pr(X(A^t,Z^t)\geq Z^t)\right). \tag{B37}$$

By Proposition 2, the optimal solution for a provider of type t is $(A^t_{\mathrm{FB}}, Z^t_{\mathrm{FB}})$. Therefore, we have proved that the first-best outcome is achieved by the nonlinear performance-based contract. □

Proof of Proposition 8

First, note that

$$\begin{aligned} F^H &= o^H E_{D_0}\left[(D_0 - C + A^H_{\mathrm{FB}})^+\right] + b\lambda(1-\theta)Pr(X(A^H_{\mathrm{FB}}, Z^H_{\mathrm{FB}}) \geq Z^H_{\mathrm{FB}}) \\ &> o^L E_{D_0}\left[(D_0 - C + A^H_{\mathrm{FB}})^+\right] + b\lambda(1-\theta)Pr(X(A^H_{\mathrm{FB}}, Z^H_{\mathrm{FB}}) \geq Z^H_{\mathrm{FB}}) \\ &\geq o^L E_{D_0}\left[(D_0 - C + A^L_{\mathrm{FB}})^+\right] + b\lambda(1-\theta)Pr(X(A^L_{\mathrm{FB}}, Z^L_{\mathrm{FB}}) \geq Z^L_{\mathrm{FB}}) = F^L. \end{aligned} \tag{B38}$$

Below we will prove that for any contract (F,K) satisfying conditions

$$F \geq F^H, \tag{B39}$$

and

$$F - K \leq o^L E_{D_0}[(D_0 - C + \theta\lambda)^+], \tag{B40}$$

a provider of type t will always choose A^t and Z^t such that the service-level constraint is not violated.

Indeed, if a provider of type t chooses A^t and Z^t that violate the service-level constraint, then the condition

$$F - K \leq o^L E_{D_0}[(D_0 - C + \theta\lambda)^+] \tag{B41}$$

implies that the objective function value for the provider is

$$F^t - K - o^t E_{D_0}[(D_0 - C + A^t)^+] - b\lambda(1-\theta)Pr(X(A^t,Z^t)\geq Z^t), \tag{B42}$$

which is smaller than

$$o^t E_{D_0}[(D_0 - C + \theta\lambda)^+] - o^t E_{D_0}[(D_0 - C + A^t)^+] - b\lambda(1-\theta)Pr(X(A^t, Z^t) \geq Z^t). \tag{B43}$$

The latter is a negative number because $A^t \geq \theta\lambda$, which is a necessary condition to ensure the stability of the queueing system. On the other hand, if a provider of type t chooses $A^t = A^t_{\text{FB}}$ and $Z^t = Z^t_{\text{FB}}$, then the service-level constraint is not violated, the transfer payment is F^t, and the objective function value at $(A^t_{\text{FB}}, Z^t_{\text{FB}})$ for the provider is $F - F^t \geq 0$. Therefore, for a provider of type t, $(A^t_{\text{FB}}, Z^t_{\text{FB}})$ is a better solution than any solution (A^t, Z^t) that violates the service-level constraint.

The problem for a provider of type-t is equivalent to

$$\max_{(A^t, Z^t) \in \mathcal{R}(M,C,\theta,\lambda)} \left(F^t - o^t E_{D_0}[(D_0 - C + A^t)^+] - b\lambda(1-\theta)Pr(X(A^t, Z^t) \geq Z^t)\right). \tag{B44}$$

Since F^t is constant, the provider's problem is further equivalent to

$$\min_{(A^t, Z^t) \in \mathcal{R}(M,C,\theta,\lambda)} \left(o^t E_{D_0}[(D_0 - C + A^t)^+] + b\lambda(1-\theta)Pr(X(A^t, Z^t) \geq Z^t)\right). \tag{B45}$$

By Proposition 2, the optimal solution for a provider of type t is $(A^t_{\text{FB}}, Z^t_{\text{FB}})$.

So far we have proved that if $F \geq \underline{F}$ and $F - K \leq o^L E_{D_0}[(D_0 - C + \theta\lambda)^+]$, then the optimal solution for a provider of type-t is $(A^t_{\text{FB}}, Z^t_{\text{FB}})$, the participation constraint is satisfied, and the objective function value for the purchaser is F. Since the purchaser wants to minimize her transfer payment to the provider, the minimum transfer payment is $\underline{F}$. On the other hand, the above procedure has shown that $F = \underline{F}$ and $K = \underline{K}$ is a feasible solution to the purchaser with a transfer payment of $\underline{F}$. Furthermore, we prove that any contract (F, K) such that $F < \underline{F}$ is infeasible because $F - K \leq F < \underline{F}$, which shows that the participation constraint is violated. Therefore, ($\underline{F}$, $\underline{K}$) is an optimal threshold-penalty PBC contract for the purchaser. The results in parts (b) and (c) of this Proposition follow immediately. □