MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 14, No. 4, Fall 2012, pp. 654–669
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

http://dx.doi.org/10.1287/msom.1120.0402

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, Pennsylvania 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 United States (under the name of "pay for performance") and abroad ("payment by results" in the United Kingdom). 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. We address the optimal contracting problem in a principal–agent framework where a service purchaser (the principal) minimizes her cost of purchasing the services and achieves the performance target (a waiting-time target) while taking into account the response of the provider (the agent) to the contract terms. Given the incentives offered by the contract, the provider maximizes his payoff by allocating his outpatient service capacity among three patient groups: urgent patients, dedicated advance patients, and flexible advance patients. We model the appointment dynamics as that of an $M/D/1$ queue and analyze several contracting approaches under adverse selection (asymmetric information) and moral hazard (private actions) settings. Our results show that simple and popular schemes used in practice cannot implement the first-best solution and that the linear performance-based contract cannot implement the second-best solution. 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 advance patients are dedicated.

Key words: healthcare; performance-based contracting; principal–agent theory; queueing theory
History: Received: September 27, 2011; accepted: March 18, 2012. Published online in *Articles in Advance* August 3, 2012.

1. Introduction

As the U.S. healthcare system is preparing to face a set of fundamental changes, the 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 2007, McGlynn et al. 2003) suggests that reform of the overall healthcare system should include changes to the mechanisms of compensating healthcare service providers. In the domain of publicly financed healthcare programs, Medicare, which leads both in terms of the number of patients covered and in terms of financial spending, is currently using the fee-for-service (FFS) scheme of physician compensation that arguably encourages providers to increase the service volume and to focus on more expensive treatment options. In addition, because FFS payments are not tied to the quality of provided services as measured by patient experiences and clinical outcomes, there exist no incentives for preventive activities or for coordination of patient care. The limitations of the FFS approach are summarized in the seminal Institute of Medicine (2007) report, which calls for the introduction of an alternative, "pay-for-performance" (P4P) provider compensation scheme. Under the P4P scheme, not only the quantity but also the quality of provided services influences the compensation amounts. In all cases of P4P adoption, the reported quality metrics include prophylaxis measures as well as clinical outcomes (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 2011).

Although the P4P framework is only now emerging from its pilot status in the United States, 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 by the National Health Service (NHS), which coordinates both the financing and the delivery of healthcare services. Since 2002, the NHS has been using a system of hospital financing called "payment by results," or PbR (since 2004 this system has also been applied to primary care physicians).[1] Similar to the FFS 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. Under the PbR system, primary care trusts (the commissioning agencies of the NHS) are free to purchase healthcare services from any qualified local provider. Unlike the fee-for-service approach, PbR considers 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[2] (most specialist care in the United Kingdom is done in state-managed hospitals). A representative example of how patient waiting times influence provider compensation is given by the 2008 standard NHS contract for acute services (UK Department of Health 2012), which stipulates penalties of up to 5% of the revenue from elective services for violating the 18-week waiting target. Recently, to facilitate better patient access to care and to streamline the management of outpatient appointments, a nationwide electronic appointment booking system, Choose and Book (CaB), was set up.

Although 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 advance 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). Second, 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 its appointment waiting list. The revenue amount and the access-to-care penalty value form the core of the hospital's performance-based contract 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, whereas the government knows only the distribution of its potential values.

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. Using this principal–agent framework, we analyze both the FFS and PBC 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 the performance of a simple contract that applies the same parameters to all agents, irrespective of their overtime cost values. In our analysis, we gain important insights by

[1] See http://www.dh.gov.uk/health/category/policy-areas/nhs/resources-for-managers/payment-by-results/ (last accessed July 21, 2012).

[2] See http://www.nhs.uk/choiceintheNHS/Rightsandpledges/Waitingtimes/Pages/Guide_to_waiting_times.aspx (last accessed July 21, 2012).

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 linear PBC approaches produce the same outcome, irrespective of whether the information setting is symmetric or asymmetric. However, if the agent's decisions are not observable and contractible, linear 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 allocation 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 first-best solution (under symmetric information) and second-best solution (under asymmetric information), respectively, as well as their implementations when the provider has private actions. In §6, we consider threshold penalty performance-based contracts, which can achieve the first-best outcome for any diverting rate, and which can also achieve the second-best outcome for the special case of dedicated-only patients. A case study is presented in §7. We conclude the paper in §8.

2. Literature Review

Goddard et al. (2000) and Farrar et al. (2007) described 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. De Fraja (2000) underscored the information asymmetry between a purchaser of services (government agency) and a service provider (hospital) inherent in healthcare settings and presented a stylized model of FFS contracting based on the principal-and-agent framework. 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 that are closely related to our work.

In the call-center context, Ren and Zhou (2008) and Hasija et al. (2008) studied coordination mechanisms in the setting where a client company outsources call-center operations to a vendor. Ren and Zhou (2008) modeled call-center operations using a fluid approximation to a $G/G/s$ queue with customer abandonment. Hasija et al. (2008) modeled a call center as an $M/M/N$ queue with customer abandonment and used a diffusion approximation. 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 research agenda, our modeling approach differs from the ones adopted by Ren and Zhou (2008) and Hasija et al. (2008) in several essential ways, reflecting the reality of a typical outpatient setting. First, our model explicitly treats outpatient appointment and service dynamics as those of an $M/D/1$ queue without using first-moment or diffusion approximations. Second, 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) considered 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, whereas in our model the information rent is unavoidable. Third, 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 a rather different task of allocating fixed service capacity among three different patient groups.

In the context of the after-sales service supply chains for multicomponent products, Kim et al. (2007) introduced a multitask principal–agent model to analyze contracts observed in practice. An important difference between our work and that of 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, whereas in the paper by Kim et al. (2007), the same information is available to the risk-averse principal and agents.

The number of applications of contract theory to healthcare services, although somewhat limited compared to retail and other service supply chains, has been growing in recent years, in part because of the increased popularity of the performance-based contracts. Lu and Donaldson (2000) presented a review of the economics literature dealing with performance-based contracting and underscored the inherent informational advantage that healthcare providers have over patients as well as purchasing agencies as one of the major sources of potential market failure in the healthcare domain.[3] Under a dynamic principal-agent

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

framework, Fuloria and Zenios (2001) studied an outcome-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 by 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, whereas Fuloria and Zenios (2001) considered only the moral hazard setting, we analyze an information asymmetry setting that leads to both moral hazard and adverse selection. Finally, Fuloria and Zenios (2001) focused on the linear contract structure, whereas we also study nonlinear, threshold performance-based contracts. Lee and Zenios (2012) studied evidence-based incentive systems within a multitask principal–agent model in the context of dialysis treatment for patients with end-stage renal disease. So and Tang (2000) considered a Medicare contract for the reimbursement of drug prescriptions in an outpatient environment with a clinical outcome-based performance metric and derived 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, because of the context of the problem analyzed, 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). Although 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.

Our work is also related to the appointment scheduling literature, which focuses on the optimal appointment capacity allocation policies in the presence of patient choice (Gupta and Wang 2008), no-shows (Liu et al. 2010), and multiple patient priorities (Patrick et al. 2008). The most important distinctive feature of our work is that we embed an appointment capacity management problem into a strategic contracting interaction between a service provider and a service purchaser.

In our model, we use a principal–agent setup 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 contracting problem in which a purchaser of services (a government agency, such as a primary care trust in the United Kingdom) offers a contract to a provider (a hospital) to deliver outpatient services. In particular, our model is designed to describe a nonsurgical outpatient specialist care environment (such as cardiology, neurology, etc.) found in many UK hospitals, where a group of physicians work in the same clinic center.

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 that can be served either on the current day or on a future date, and same-day appointments that 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.

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 system if she finds an available time slot or through the override phone-based system (in the United Kingdom, the Telephone Appointment Line) if no time slot is available from her chosen provider through the CaB system. In other words, a dedicated patient insists on being serviced by her first-choice provider for reasons of geographical proximity, the 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 system 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 system with the first-choice provider turns out to be a dedicated patient with probability θ, a parameter that 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, whereas $\theta = 0$ would characterize an undistinguished facility with easy-to-find substitutes. With outpatient appointments scheduled using the CaB system, patients who are not able to schedule an appointment at the hospital of their choice only observe that there are no appointments available at all—in particular, they are not able to ascertain the appointment capacity allocation policy used by the hospital.

In reality, patients may make choices not only in terms of 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 system. 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 cumulative distribution function (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. Although every hospital in the United Kingdom is required to manage its advance appointments using the CaB system, the exact fraction of its service capacity to be released to the CaB system is within 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.

The (A, Z) policy we use in our model closely reflects the actual appointment management practices used by the NHS hospitals. In their use of the CaB system, NHS hospitals control their appointment capacity by introducing the available appointment horizon, often referred to as the "polling range" (Z), and restrict the daily number of appointment slots released to the online system (A). Hospital's Choose and Book managers are typically responsible for the job of allocating the appointment slots to the CaB system. All the required information on the past and the current patient appointments is readily available from the hospital database; however, the values of A and Z are currently set on an ad hoc basis.

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. This assumption 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. The single-server assumption we use to describe the evolution of appointment backlog in a hospital with multiple physicians serving patients in parallel requires a justification. Let us consider a hospital specialty clinic that has P physicians, each having daily service capacity measured by S appointment slots. On a daily basis, the total appointment capacity for such a clinic is therefore equal to PS slots. Suppose that daily demand for appointments can be described as a stationary Poisson process with rate λ. Because our analysis is not focused on minute-by-minute details of patient service but rather on daily evolution of the appointment backlog, we could use a discrete-time approach to describing the appointment dynamics, with "day" being the discrete time unit. Typical appointment backlogs in outpatient care stretch for weeks, and under this approach, the daily change in the actual state of the backlog is equal to the difference between the total daily demand for appointments and the total daily supply of appointment slots, PS. Note, however, that this daily change is equivalent to the one obtained from a continuous-time $M/D/1$ model, with a single server operating for a day at the rate of $A = PS$ slots per day, except on those rare days when the appointment backlog is smaller than the daily appointment capacity. An alternative description of such appointment dynamics process would be achieved by using a continuous-time $M/D/P$ model operating at the rate of S slots per server per day. Among these three modeling choices (the discrete-time daily model and $M/D/1$ and $M/D/P$ queues), we have selected $M/D/1$ dynamics on the basis of its relative tractability, which allowed us to obtain partial characterization of the appointment backlog performance measures.

Note that the patient appointment backlog grows both during and outside of office hours, because 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. Although such a dynamic 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 the rate of $\theta\lambda/A$ per slot if the appointment backlog is equal to or larger than Z (when no slots are available through the CaB system, only dedicated patients can get appointments). Note that this queueing system, which we denote as the *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 $\lambda(1-(1-\theta)\cdot \Pr(X(A, Z) \geq Z))$. Note that because 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. Our analysis assumes that a patient turned away by a hospital of his/her choice will be served at another hospital. Although such patient transfers do not directly impact social welfare, they are often undesirable because they come in to contradiction with hospitals' implicit obligation to serve the local population. As a result, hospitals are wary of the patient transfer process resulting from finite capacity horizon Z used in the online appointment system. To model hospital aversion to patient transfer, we include a penalty term in hospital objective function. If b is the cost for diverting one patient, the expected daily diverting cost is given by $P(A, Z) = b\lambda(1-\theta)\Pr(X(A, Z) \geq Z)$.

Information asymmetries between purchasers and providers and between providers and clients pervade healthcare service supply chains (Arrow 1963, De Fraja 2000, Haas-Wilson 2001, Bloom et al. 2008, Levaggi and Levaggi 2010). In particular, the cost structure of the service provider often remains private knowledge that is neither fully communicated to nor fully verified by the purchaser. In our analysis, we assume that the value of the overtime cost constitutes private information that reflects the provider's ability to stretch its daily service capacity to match unexpected surges in same-day patient demand. More specifically, we assume that the overtime cost per patient could take one of the k values, o^t, $t \in \{1, \dots, k\}$, with $o^1 < \cdots < o^k$. 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

$$O^t(A) = o^t E_{D_0}[(D_0 - C + A)^+], \tag{1}$$

where D_0 is a random variable representing the same-day demand. Under information asymmetry with respect to the overtime cost, the purchaser of services only knows the distributional information regarding the value of the provider's overtime cost: for the purchaser, the hospital belongs to type t with probability p^t, $0 < p^t < 1$, so that $\sum_{t=1}^{k} p^t = 1$.

3.2. Performance Metric: Patient Waiting Time

The set of measures used in practice for evaluating the performance of healthcare providers includes 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. Because 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 $L_q(A, Z) = \lambda(1-(1-\theta)\Pr(X(A, Z) \geq Z)) \cdot W_q(A, Z)/A$. 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.

All proofs are presented in the electronic companion (available at http://dx.doi.org/10.1287/msom.1120.0402).

Most of the results of Proposition 1 are intuitive. On one hand, an increase in θ, 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 because the left-hand side of the inequality $X(A,Z) \geq Z$ is stochastically increasing in Z.

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 $L_q(A,Z) = \rho^2/(2(1-\rho)) = \lambda^2/(2A(A-\lambda))$ and $W_q(A,Z) = \rho/(2(1-\rho)) = \lambda/(2(A-\lambda))$. Note that because the daily number of advance appointment slots is A, the expected number of *days* a patient has to wait is

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

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

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

$$\frac{W_q(A,Z)}{A} \leq M, \tag{3}$$

where M is the waiting-time target measured in days. Note that for $\theta = 1$ or $Z = \infty$, $W_q(A,Z)/A = \lambda/(2A(A-\lambda))$. From the result of Proposition 1 it follows that the service-level constraint (3) implies a lower bound for the value of A,

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

To ensure the feasibility of the capacity management problem, we require that the overall daily service capacity C is not lower than A^*:

$$C \geq A^* = \frac{\lambda}{2} + \sqrt{\frac{\lambda^2}{4} + \frac{\lambda}{2M}}. \tag{5}$$

We conclude this subsection by stating a connection between the assumption (5) and the service-level constraint (3).

Lemma 1. *Consider the modified* $M/D/1$ *queue. For any* $\theta \in [0,1]$ *and any* $Z \geq 0$, *the service-level constraint* (3) *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 ($t = 1, \ldots, k$), the expected profit is obtained by combining the transfer payment T^t with the patient-diverting and overtime costs:

$$\Pi_a^t(A^t, Z^t) = T^t - o^t E_{D_0}[(D_0 - C + A^t)^+] - b\lambda(1-\theta)\Pr(X(A^t, Z^t) \geq Z^t). \tag{6}$$

Notice that at this stage, the structure of the transfer payment term T^t remains undefined, because it may take different forms based on the type of contract being considered. The purchaser minimizes the expected cost $\Pi_p = \sum_{t=1}^k p^t T^t$, while ensuring that the patient waiting-time target (3) is met.

A linear 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 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.

We assume that the FFS payment is the same for both advance and same-day patients because these patients require similar outpatients services in a particular outpatients clinic. As the expected number of patients treated each day is equal to $\lambda_0 + \lambda(1 - (1-\theta)\Pr(X(A^t, Z^t) \geq Z^t))$, the expected daily FFS payment is $r^t(\lambda_0 + \lambda(1 - (1-\theta)\Pr(X(A^t, Z^t) \geq Z^t)))$. 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

$$\begin{aligned} T^t(r^t, l^t, A^t, Z^t) &= r^t\big(\lambda_0 + \lambda(1 - (1-\theta)\Pr(X(A^t, Z^t) \geq Z^t))\big) \\ &\quad - l^t \frac{W_q(A^t, Z^t)}{A^t}, \end{aligned} \tag{7}$$

where $\lambda_0 = E[D_0]$ is the expected value of the same-day demand. In the healthcare economics literature, it is often assumed that the service provider is altruistic and derives additional, nonmonetary 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 the expected profit. It is interesting to note that even in the United Kingdom, where healthcare providers are publicly funded nonprofit organizations, NHS foundation trusts are often described as profit maximizers (De Fraja 2000, 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 $t\in\{1,\ldots,k\}$ 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 k 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

Under the symmetric information setting, the purchaser learns the provider's type t before deciding on the contract terms, and can therefore tailor a contract to the specific provider type. 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 United States, capacity allocation policies often constitute the provider's "private actions," which remain unobservable to the purchaser. In such environments, the purchaser 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: First-Best Solution

If the provider's capacity allocation policy (A,Z) is observable and contractible, the purchaser solves the following problem faced with the provider of type $t\in\{1,\ldots,k\}$:

$$\min_{T^t,A^t,Z^t}\quad T^t(A^t,Z^t) \tag{8}$$

$$\text{s.t.}\quad (A^t,Z^t)\in\mathscr{R}(M,C,\theta,\lambda), \tag{9}$$

$$\Pi_a^t(T^t,A^t,Z^t)=T^t-o^tE_{D_0}[(D_0-C+A^t)^+]$$
$$\qquad -b\lambda(1-\theta)\Pr(X(A^t,Z^t)\ge Z^t)\ge 0, \tag{10}$$

$$T^t\ge 0, \tag{11}$$

where

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

The objective for the purchaser is to minimize the payout T^t for each type of provider. The first constraint, (9), 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 $\theta\lambda/2+\sqrt{(\theta^2\lambda^2)/4+(\theta\lambda)/(2M)}$, the value that guarantees that the expected waiting-time target is met even for $Z^t=0$. The second constraint, (10), is the individual rationality constraint, which guarantees the provider of type t accepts his designated contract. The following proposition describes the optimal solution for the complete information problem, also known as the first-best solution.

Proposition 2. (a) *For*

$$\frac{\theta\lambda}{2}+\sqrt{\frac{\theta^2\lambda^2}{4}+\frac{\theta\lambda}{2M}}\le A\le C,$$

let

$$Z_M(A)=\max_{Z\in\mathcal{N}}\left\{Z\,\middle|\,\frac{W_q(A,Z)}{A}\le M\right\}. \tag{12}$$

The family of optimal first-best contracts $(T_{\text{FB}}^t, A_{\text{FB}}^t, Z_{\text{FB}}^t)$ *is characterized by*

$$A_{\text{FB}}^t=\mathop{\arg\min}_{(\theta\lambda)/2+\sqrt{(\theta^2\lambda^2)/4+(\theta\lambda)/(2M)}\le A^t\le C}\left\{o^tE_{D_0}[(D_0-C+A^t)^+]\right.$$
$$\left.+b\lambda(1-\theta)\Pr(X(A^t,Z_M^t(A^t))\ge Z_M^t(A^t))\right\}, \tag{13}$$

$$Z_{\text{FB}}^t=Z_M(A_{\text{FB}}^t),$$

and

$$T_{\text{FB}}^t=o^tE_{D_0}[(D_0-C+A_{\text{FB}}^t)^+]$$
$$\qquad +b\lambda(1-\theta)\Pr(X(A_{\text{FB}}^t,Z_{\text{FB}}^t)\ge Z_{\text{FB}}^t).$$

(b) *The first-best capacity allocation decisions* A_{FB}^t *and* Z_{FB}^t *are nonincreasing functions of* o^t *and nondecreasing functions of* b.

The results of Proposition 2(a) 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 that the expected waiting time is as close as possible to the target value. The optimal payment T^t is set to extract the entire surplus from provider of type t, so that $\Pi^t_a(T^t_{\text{FB}}, A^t_{\text{FB}}, Z^t_{\text{FB}}) = 0$.

It is easy to show that the linear performance-based contract defined in (7) can achieve the optimal first-best performance if and only if

$$\begin{aligned} r^t_{\text{FB}} = {} & \big(o^t E_{D_0}[(D_0 - C + A^t_{\text{FB}})^+] \\ & b\lambda(1-\theta)\Pr(X(A^t_{\text{FB}}, Z^t_{\text{FB}}) \ge Z^t_{\text{FB}}) \\ & + l^t_{\text{FB}} W_q(A^t_{\text{FB}}, Z^t_{\text{FB}})/A^t_{\text{FB}}\big) \\ & \cdot \big(\lambda_0 + \lambda(1 - (1-\theta)\Pr(X(A^t_{\text{FB}}, Z^t_{\text{FB}}) \ge Z^t_{\text{FB}}))\big)^{-1}, \\ & l^t_{\text{FB}} \in \mathscr{R}^+,\ t \in \{1, \ldots, k\}. \end{aligned} \tag{14}$$

As (14) 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. 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* $o^t = 0, t \in \{1, \ldots, k\}$, *the first-best solution is given by* $A^t_{\text{FB}} = C, Z^t_{\text{FB}} = \infty$.

(b) *For* $b = 0$, *the first-best solution is given by* $A^t_{\text{FB}} = (\theta\lambda)/2 + \sqrt{(\theta^2\lambda^2)/4 + (\theta\lambda)/(2M)}, Z^t_{\text{FB}} = 0, t \in \{1, \ldots, k\}$.

(c) *For* $\theta = 1$, *the first-best solution is given by* $A^t_{\text{FB}} = A^* = \lambda/2 + \sqrt{\lambda^2/4 + \lambda/(2)M}, Z^t_{\text{FB}} \in \mathcal{N}, t \in \{1, \ldots, k\}$.

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 ($A = \theta\lambda/2 + \sqrt{(\theta^2\lambda^2)/4 + (\theta\lambda)/(2M)}$ 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 can be expressed as $T_{\text{FB}} = \sum_{t=1}^k p^t T^t_{\text{FB}} = (\sum_{t=1}^k p^t o^t) E_{D_0}[(D_0 - C + A^*)^+]$.

4.2. Private Actions: Implementing the First-Best Outcome

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, 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, and PbR (Mannion et al. 2008).

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. 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). It is easy to show that neither block nor FFS contracts can achieve the first-best outcome.

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 (7). The following proposition identifies linear performance-based contract parameters that achieve the first-best outcome for the setting in which all patients are dedicated.

Proposition 3. *Let* $\lambda_0 = E_{D_0}[D_0]$. *For* $\theta = 1$, *the first-best outcome is obtained by*

$$\tilde{r}^t = \frac{o^t}{\lambda + \lambda_0}\left(\frac{\lambda(1 - F_{D_0}(C - \lambda/2 - \sqrt{\lambda^2/4 + \lambda/(2M)}))}{4M\sqrt{\lambda^2/4 + \lambda/(2M)}} + E_{D_0}\left[\left(D_0 - C + \frac{\lambda}{2} + \sqrt{\frac{\lambda^2}{4} + \frac{\lambda}{2M}}\right)^+\right]\right), \tag{15}$$

$$\tilde{l}^t = o^t\left(1 - F_{D_0}\left(C - \frac{\lambda}{2} - \sqrt{\frac{\lambda^2}{4} + \frac{\lambda}{2M}}\right)\right) \Big/ \left(4M^2\sqrt{\frac{\lambda^2}{4} + \frac{\lambda}{2M}}\right). \tag{16}$$

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 after Proposition 2, but are uniquely determined by (15) and (16). 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. Asymmetric Information

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: Second-Best Solution

The information asymmetry in assessing the provider's overtime costs leads to the adverse selection problem (Bolton and Dewatripont 2005), and the purchaser must design a contract menu applying the revelation principle. More specifically, let Π_a^{ts} 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):

$$\Pi_a^{ts}(T^s, A^s, Z^s) = T^s(A^s, Z^s) - o^t E_{D_0}[(D_0 - C + A^s)^+] - b\lambda(1-\theta)\Pr(X(A^s, Z^s) \geq Z^s). \quad (17)$$

Note that $\Pi_a^t(T^t, A^t, Z^t)$ defined in (6) is equivalent to $\Pi_a^{tt}(T^t, A^t, Z^t)$. The purchaser's problem can be formulated as follows:

$$\min_{T^t, A^t, Z^t} \sum_{t=1}^{k} p^t T^t(A^t, Z^t) \quad (18)$$

$$\text{s.t.} \quad (A^t, Z^t) \in \mathcal{R}(M, C, \theta, \lambda), \quad t = 1, \ldots, k, \quad (19)$$

$$\Pi_a^{tt}(T^t, A^t, Z^t) \geq 0, \quad t = 1, \ldots, k, \quad (20)$$

$$\Pi_a^{tt}(T^t, A^t, Z^t) \geq \Pi_a^{ts}(T^s, A^s, Z^s), \quad t, s = 1, \ldots, k, \; s \neq t, \quad (21)$$

$$T^t \geq 0, \quad t = 1, \ldots, k. \quad (22)$$

The waiting-time target and stability constraints (19) and the individual rationality constraints (20) are the analogues of the constraints (9) and (10) in the symmetric information setting. Constraints (21) are the incentive compatibility constraints. 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 (7). The following proposition characterizes the structure of the second-best solution.

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

$$A_{\text{SB}}^t = \operatorname*{arg\,min}_{(\theta\lambda)/2 + \sqrt{(\theta^2\lambda^2)/4 + (\theta\lambda)/(2M)} \leq A^t \leq C} \left\{ \hat{o}^t E_{D_0}[(D_0 - C + A^t)^+] + b\lambda(1-\theta)\Pr(X(A^t, Z_M(A^t)) \geq Z_M(A^t)) \right\}, \quad (23)$$

$$Z_{\text{SB}}^t = Z_M(A_{\text{SB}}^t),$$

where $\hat{o}^1 = o^1$, $\hat{o}^t = o^t + \sum_{s=1}^{t} p^s/p^t(o^t - o^{t-1})$, $t = 2, \ldots, k$, *and* $Z_M(A) = \max_{Z \in \mathcal{N}}\{Z \mid W_q(A, Z)/A \leq M\}$.

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

$$T_{\text{SB}}^k = o^k E_{D_0}[(D_0 - C + A_{\text{SB}}^k)^+] + b\lambda(1-\theta)\Pr(X(A_{\text{SB}}^k, Z_{\text{SB}}^k) \geq Z_{\text{SB}}^k)$$

and

$$T_{\text{SB}}^t = o^t E_{D_0}[(D_0 - C + A_{\text{SB}}^t)^+] + b\lambda(1-\theta)\Pr(X(A_{\text{SB}}^t, Z_{\text{SB}}^t) \geq Z_{\text{SB}}^t) + \sum_{s=t}^{k-1}(o^{s+1} - o^s)E_{D_0}[(D_0 - C + A_{\text{SB}}^{s+1})^+], \quad t = 1, \ldots, k-1.$$

(c) *Let A_{FB}^t and Z_{FB}^t be the first-best capacity allocation controls defined in* (13). *Then, $A_{SB}^1 \geq \cdots \geq A_{SB}^k$, $A_{SB}^t \leq A_{FB}^t$, $t = 1, \ldots, k$, and $Z_{SB}^1 \geq \cdots \geq Z_{SB}^k$, $Z_{SB}^t \leq Z_{FB}^t$, $t = 1, \ldots, k$.*

Proposition 4(a) shows that the second-best capacity allocation policy is obtained by solving k separate optimization problems, one for each provider type, with a structure identical to that of the first-best problem (13). The optimization problem for the lowest-cost provider is identical to (13), whereas those for higher-cost providers use the value of the overtime cost for the corresponding provider type adjusted upward due to the presence of information asymmetry. Part (b) shows that the payout to the highest-cost provider is still equal to its operational cost (equal to the sum of the overtime cost and the patient diverting cost), and the payouts to lower-cost providers are higher than their operational costs: an additional information rent, $\sum_{s=t}^{k-1}(o^{s+1} - o^s)E_{D_0}[(D_0 - C + A_{\text{SB}}^{s+1})^+]$, is paid to each lower-cost provider as a result of the existing information asymmetry. Part (c) shows that whereas the second-best capacity allocation policy intended for the lowest-cost provider replicates the first-best policy for this provider type, the second-best capacity allocation policy intended for the higher-cost providers differs from the corresponding first-best solution. In particular, the daily capacity allocated to advance appointments under the second-best solution (A^t_{SB}) is, in general, lower than that under the corresponding first-best solution (A^t_{FB})—inefficiency created by information asymmetry.

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 ($o^t = 0, t = 1, \ldots, k$, 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.

5.2. Performance-Based Contracts with Private Actions

We next show that, in general, the second-best solution cannot be implemented under information asymmetry using linear PBC in the presence of private actions. To this end, it suffices to address 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 performance-based contracts, $\{(r^t, l^t), t = 1, \ldots, k\}$, the type-$t$ provider who reports to be of type s solves the following optimization problem:

$$\max_{\theta\lambda \le A^{ts} \le C} \big\{\Pi^{ts}_a(r^s, l^s, A^{ts}) \equiv T^s(r^s, l^s, A^{ts}) - o^t E_{D_0}[(D_0 - C + A^{ts})^+]\big\}, \quad s \in \{1, \ldots, k\}, \tag{24}$$

where $T^s(r^s, l^s, A^{ts}) = r^s(\lambda + \lambda_0) - l^s(\lambda/(2A^{ts}(A^{ts} - \lambda)))$ 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^{ts}_{PA}. 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 $\{(r^t, l^t), t = 1, \ldots, k\}$, the following hold*:

(a) *For any $t < s$, $A^{ts}_{\mathrm{PA}} \ge A^{ss}_{\mathrm{PA}}$. In particular, if $\lambda < A^{ss}_{\mathrm{PA}} < C$, then $A^{ts}_{\mathrm{PA}} > A^{ss}_{\mathrm{PA}}$.*

(b) *For any $t < s$, $A^{tt}_{\mathrm{PA}} \ge A^{st}_{\mathrm{st}}$. In particular, if $\lambda < A^{st}_{\mathrm{PA}} < C$, then $A^{tt}_{\mathrm{PA}} > A^{st}_{\mathrm{PA}}$.*

(c) *For any $t \ne s$, A^{ts}_{PA} is increasing in l^s. In particular,*

$$A^{ts}_{\mathrm{PA}} \ge A^* = \frac{\lambda}{2} + \sqrt{\frac{\lambda^2}{4} + \frac{\lambda}{2M}}$$

if and only if $l^s \ge \tilde{l}^t$, where

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

Proposition 5 shows that given any menu of contracts, a higher-cost provider does not choose a higher capacity level than a lower-cost provider. In addition, it states that a provider's capacity allocated to advance appointments, as expected, is increasing in the waiting-time penalty cost. Considering providers' optimal responses, the purchaser solves the following problem:

$$\min_{r^t, l^t} \sum_{t=1}^{k} p^t T^t(r^t, l^t, A^{tt}_{\mathrm{PA}}) \tag{25}$$

$$\text{s.t.} \quad A^{tt}_{\mathrm{PA}} \ge A^*, \quad t = 1, \ldots, k, \tag{26}$$

$$\Pi^{tt}_a(r^t, l^t, A^{tt}_{\mathrm{PA}}) \ge 0, \quad t = 1, \ldots, k, \tag{27}$$

$$\Pi^{ts}_a(r^t, l^t, A^{tt}_{\mathrm{PA}}) \ge \Pi^{ts}_a(r^s, l^s, A^{ts}_{\mathrm{PA}}), \quad t, s = 1, \ldots, k,\ t \ne s, \tag{28}$$

$$r^t \ge 0, l^t \ge 0, \quad t = 1, \ldots, k. \tag{29}$$

A partial characterization of the optimal contract parameters in (25)–(29) is provided below.

Proposition 6. *When $\theta = 1$, under the optimal contract, $(r^t_{\mathrm{PA}}, l^t_{\mathrm{PA}})$, $t = 1, \ldots, k$, we have*

(a) *$l^t_{\mathrm{PA}} \ge \tilde{l}^t$, $t = 1, \ldots, k$, $A^{tt}_{\mathrm{PA}} > A^*$, $t = 1, \ldots, k - 1$;*

(b) *$\Pi^{kk}_a(r^k, l^k, A^{kk}_{\mathrm{PA}}) = 0$, $\Pi^{tt}_a(r^t, l^t, A^{tt}_{\mathrm{PA}}) > (o^k - o^t) \cdot E_{D_0}[(D_0 - C + A^*)^+]$, $t = 1, \ldots, k - 1$;*

(c) *$\sum_{t=1}^{k} p^t T^t(r^t, l^t, A^{tt}_{\mathrm{PA}}) > o^k E_{D_0}[(D_0 - C + A^*)^+]$.*

Proposition 6 shows that compared to the first-best (and the second-best) capacity allocation, A^*, providers of all types tend to allocate higher capacities to serving advance appointments. Parts (b) and (c) of Proposition 6 provide further evidence of linear PBC's inability to achieve the second-best outcome: the purchaser provides a higher rent to the low-cost providers 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 performance-based contract 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 performance-based contract 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 nonlinear 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, let*

$$F^t = o^t E_{D_0}[(D_0 - C + A^t_{\mathrm{FB}})^+] + b\lambda(1-\theta) \cdot \Pr(X(A^t_{\mathrm{FB}}, Z^t_{\mathrm{FB}}) \geq Z^t_{\mathrm{FB}}), \quad t = 1, \ldots, k, \tag{30}$$

and let K be the positive constant such that $K > F^t - o^t E_{D_0}[(D_0 - C + \theta\lambda)^+]$. 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 is not:

$$T^t = \begin{cases} F^t & \text{if } W_q(A^t, Z^t)/A^t \leq M, \\ F^t - K & \text{if } W_q(A^t, Z^t)/A^t > M. \end{cases} \tag{31}$$

Any threshold-penalty performance-based contract specified by (30) *and* (31) *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 performance-based contract $(\underline{F}, \underline{K})$ defined by $\underline{F} = F^k$ and $\underline{K} = F^k - o^1 E_{D_0}[(D_0 - C + \theta\lambda)^+]$, where F^k is given by* (30).

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

(b) *Under the contract $(\underline{F}, \underline{K})$, the optimal capacity allocation policy for each provider type is described by $A^t_{\mathrm{TP}}(\underline{F}, \underline{K}) = A^t_{\mathrm{FB}}, Z^t_{\mathrm{TP}}(\underline{F}, \underline{K}) = Z^t_{\mathrm{FB}}$, and the resulting expected transfer payments are $T^t(\underline{F}, \underline{K}) = \underline{F}, t = 1, \ldots, k$.*

(c) *The threshold-penalty performance-based contract $(\underline{F}, \underline{K})$ achieves the second-best solution for $\theta = 1$: $A^t_{\mathrm{TP}}(\underline{F}, \underline{K}) = A^*, T^t(\underline{F}, \underline{K}) = o^k E_{D_0}[(D_0 - C + A^*)^+]$, $t = 1, \ldots, k$.*

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 all k 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.

7. Case Study: NHS Shetland

This section provides a case study of a small NHS trust in Scotland, NHS Shetland, to illustrate the nature of the first- and and second-best solutions.

7.1. Data and Model Calibration

We use the actual outpatient appointment data collected by the National Health Service of Scotland[4] to estimate some of the parameters of our model. Table 1 summarizes the collected quarterly data on the outpatient appointments, aggregated across all clinical specialties, for the period of January 2008–December 2009. In particular, for each quarter, Table 1 reports the total number of patients served, the total number of patients waiting for their appointments at the end of the quarter, and various characteristics of the waiting-time distribution for patients served during the quarter.[5] The period 2008–2009 was a transition period during which the 18-week waiting target was introduced. As a result, the number of served patients has increased between 2008 and 2009, and the patient appointment waits were reduced. Because we are interested in calibrating the model based on stationary $M/D/1$ dynamics, our estimates reflect time-averaged system behavior over this period. The number of new appointments (total number of patients who requested service during a quarter) was calculated, for each quarter, by adding the number of patients waiting at the end of the current quarter and the number of patients seen during the current quarter, and subtracting the number of patients waiting at the end of the previous quarter.

In our parameter estimation approach, we used the average (over all quarters) values to estimate the parameters A, λ, θ, and Z. In particular, we have assumed that the average values of the wait characteristics in Table 1 represent a stationary distribution of patient appointment wait generated by a modified $M/D/1/Z$ queue, and focused on estimating the values of A, λ, Z, and θ for such a queue. First, we have estimated the daily number of appointment slots A as the average number of patients served on each workday, rounded to the nearest integer. For example, the estimate for the daily number of appointment slots is $A_S = \lceil 950/65.36 \rceil = \lceil 14.53 \rceil = 15$, where $65.36 = 5 \times 91.5/7$ is the average number of workdays in a quarter. (We rounded up the estimator 950/65.36 to maintain the integer value of capacity measured in terms of the number of appointment slots.) Parameters λ, θ, and Z were estimated by minimizing

[4] Data are available at http://www.isdscotland.org/Health-Topics/Waiting-Times/Publications/data-tables.asp? (last accessed September 27, 2011).

[5] The "wait" here refers to the interval between the time patient got an appointment and the time patient was seen by a physician. Note that the wait distribution reflects the data for all patients served during a particular quarter.

Table 1 Quarterly Outpatient Appointment Data from January 1, 2008, to December 31, 2009, for NHS Shetland

	1st quarter, 2008	2nd quarter, 2008	3rd quarter, 2008	4th quarter, 2008	1st quarter, 2009	2nd quarter, 2009	3rd quarter, 2009	4th quarter, 2009	Average
Patients served	784	834	746	913	954	1,113	1,112	1,147	950
Patients waiting	707	682	601	457	518	620	610	480	584
New appointments		809	665	769	1,015	1,215	1,102	1,017	942
Median wait (days)	49	50	58	45	37	31	32	35	42
90th percentile wait (days)	91	93	104	98	81	61	70	77	84
% wait up to 3 weeks	23	24	28	29	28	35	35	34	30
% wait up to 6 weeks	44	43	41	47	57	70	62	58	53
% wait up to 9 weeks	69	67	55	70	78	93	83	76	74
% wait up to 12 weeks	86	85	77	82	93	100	100	96	90
% wait up to 15 weeks	98	95	91	94	99	100	100	96	97
% wait up to 18 weeks	100	100	100	100	100	100	100	97	100

the sum of the squared deviations between the values of the expected arrival rate of new appointments on each workday, $\Lambda(A,\lambda,\theta,Z)=\lambda(1-\Pr(X(A,Z)\geq Z))+\lambda\theta\Pr(X(A,Z)\geq Z)$, the median stationary backlog $B_{50}(A,\lambda,\theta,Z)$, and the 90th percentile of the stationary backlog distribution $B_{90}(A,\lambda,\theta,Z)$ for a modified $M/D/1/Z$ queue, respectively, and the average values observed at this trust. In other words, our estimation is equivalent to the following minimization problem: $(\lambda_S,\theta_S,Z_S)=\arg\min_{\lambda,\theta,Z}(\|\Lambda(A_S,\lambda,\theta,Z)-14.86\|^2+\|B_{50}(A_S,\lambda,\theta,Z)-30\|^2+\|B_{90}(A_S,\lambda,\theta,Z)-60\|^2)$, where $14.86=15\times 942/950$ is the appropriately scaled-up value of the observed average rate of arrival of new appointment requests on a workday, and $30=5\times 42/7$ $(60=5\times 84/7)$ is the median (90th percentile) appointment backlog expressed in terms of the number of workdays. To control the computational time, in our search for minimizing values of θ and Z we initially limited our choice to $\theta=0,0.1,\ldots,0.9,1$ and $Z=mA$, $m\in\mathcal{N}$, and then conducted a more refined local search.

Our estimation procedure results in $\lambda_S=15$, $\theta_S=0.94$, and $Z_S=968$ appointment slots for this hospital. We use these estimates in our numerical examples in the remainder of this section. As our estimated parameters indicate, the outpatient appointment dynamics in NHS Shetland appear to be best represented as those of a completely loaded queue with the appointment capacity fully utilized by the demand for appointments, serving a nearly perfectly dedicated patient population. Both of these features are not surprising given the extensive observed appointment waits and the relatively isolated geographical location of this region.

To get an indication of the applicability of a modified $M/D/1/Z$ queue model for describing the observed appointment dynamics, it is interesting to compare the CDF values for the stationary appointment backlog distribution (percentiles of patients waiting up to a certain number of weeks) in such a queue generated using the estimated parameters and the average values reported in Table 1. Figure 1 reports such a comparison, indicating that a modified $M/D/1/Z$ queue provides, overall, an adequate approximation to the actual wait dynamics, deviating from the observed values in a nonsystematic fashion: somewhat underestimating the probability of wait for low and moderate wait durations, and overestimating it for high durations.

7.2. First-Best and Second-Best Solutions

Proposition 2 shows that in the first-best solution, the capacity allocation decisions A^t_{FB} and Z^t_{FB} are monotone in the overtime cost o^t and diversion penalty b. Our numerical results indicate that these monotone properties may also extend to diversion parameter θ and waiting-time target M. As follows from the monotonicity properties of $W_q(A,Z)/A$ 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. Figure 2 illustrates the first-best capacity allocation decisions as functions of the waiting-time target M in nine 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

Figure 1 CDF Values of the Stationary Appointment Wait Distribution: A Modified $M/D/1/Z$ Queue Generated by the Estimated Parameters vs. Average Values from Table 1

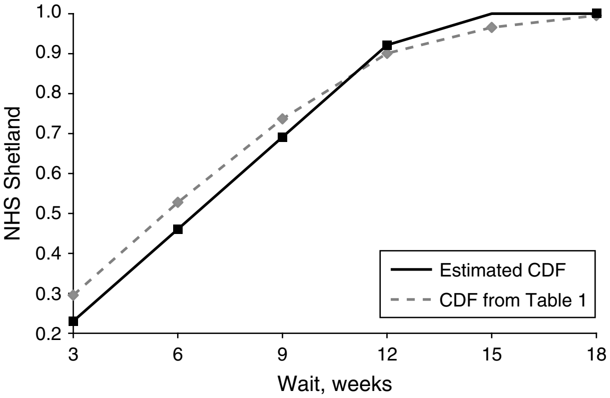
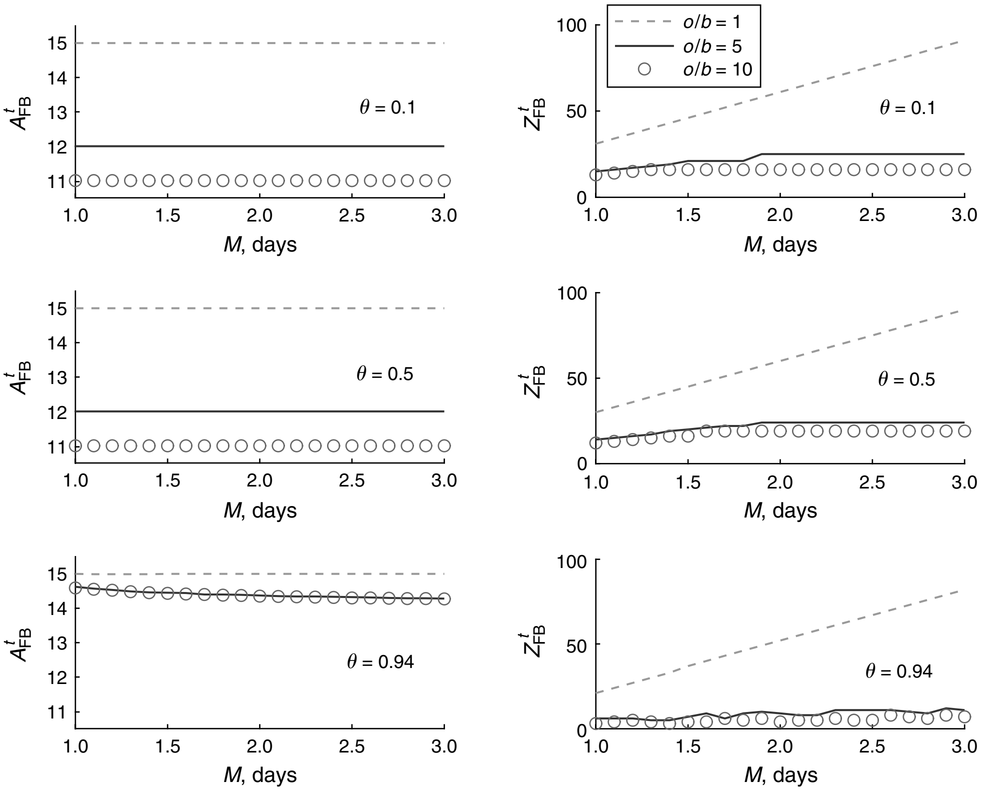


Figure 2 Optimal First-Best Capacity Allocation Decisions A^t_{FB} and Z^t_{FB} as Functions of the Waiting-Time Target M for $\lambda = 15$ (as Estimated for NHS Shetland), $\lambda_0 = 4$, and $C = 18$



$\theta = 0.94$, corresponding to mostly dedicated patients, as estimated for NHS Shetland) 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. On the other hand, as financial penalties associated with patient diversions diminish, the composition of the patient population plays an increasingly important role 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 nonmonotone fashion as functions of θ. We now demonstrate the properties of the second-best solution using the NHS Shetland data. For ease of exposition, we assume the provider can be of one of two types: high or low cost, indicated by superscripts H and L, respectively. In settings with mixed patient populations, the threshold-penalty performance-based contract may no longer be able to achieve the second-best performance. Figure 3 depicts the second-best solution (A^H_{SB}, Z^H_{SB}, A^L_{SB}, Z^L_{SB}) and the threshold-penalty PBC solution (A^H_{TP}, Z^H_{TP}, A^L_{TP}, Z^L_{TP}) in the same patient-mix settings as in Figure 2: $\theta = 0.1$, $\theta = 0.5$, and $\theta = 0.94$. The capacity allocation policies shown in Figure 3 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 $(\underline{F}, \underline{K})$ performance-based contract and in the second-best solution coincide. Third, under the $(\underline{F}, \underline{K})$ performance-based 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 performance-based contract 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 toward mostly dedicated patients.

8. Conclusions

As ever increasing numbers of healthcare organizations recognize service access as an important component of the quality of healthcare services,

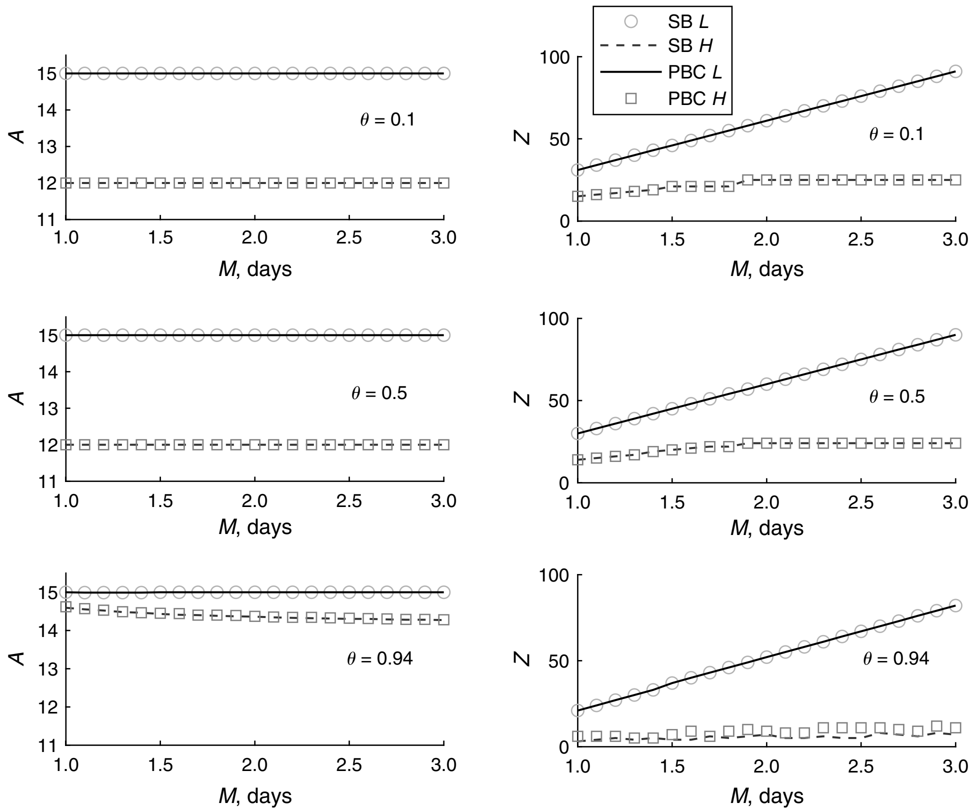


Figure 3 Optimal Second-Best Solution (A^H_{SB}, Z^H_{SB}, A^L_{SB}, Z^L_{SB}) and Optimal Solution (A^H_{TP}, Z^H_{TP}, A^L_{TP}, Z^L_{TP}) for the Threshold-Penalty Performance-Based Contract as Functions of the Waiting-Time Target M, $\lambda = 15$ (as Estimated for NHS Shetland), Poisson Same-Day Demand with Rate $\lambda_0 = 4$, $o^H/b = 5$, $o^L/b = 1$, $p = 0.5$, and $C = 18$

performance-based contracts that include access performance measures gain increasing popularity in United States and abroad. Motivated by recent reforms in health service payment mechanisms in the United Kingdom, we study a performance-based approach to contracting for outpatient services used in the United Kingdom 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, advance dedicated, and advance flexible patients.

An important feature of real-life capacity allocation decisions made by care providers is their multidimensional and dynamic nature. In the present work, we adopted a simplifying approach to modeling these decisions by assuming a two-dimensional, open-loop provider's response, which allowed us to focus on key contractual issues while capturing important capacity allocation trade-offs. We believe future research can build on our findings by incorporating more realistic features of day-to-day appointment accumulation and service dynamics.

Our analysis of first- and second-best contracting problems shows that performance-based contracts are superior to both "block" contracts and FFS contracts. However, we also established that a simple linear performance-based contract is only guaranteed to achieve coordination in the case of dedicated-only patients, while failing 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. The value of the appointment waiting target as a measure of the patient access to care is a subject of intense political debate in the United Kingdom and is likely to become a part of a similar debate in the United States once tens of millions of new patients receive medical coverage in the near future. Our analysis enables policy makers to quantify the cost of shortening patient appointment backlogs and identifies the waiting-time target as a critical factor driving system performance and contract design. One indication of the relevance of our analysis for U.S. healthcare settings is a growing number of performance-based contracts applied to purchases of healthcare services in this country. For example, similar to the United Kingdom, the State of California has recently implemented a series of waiting time targets for medical services, with the target to see a specialist set at 15 business days (*California Healthline* 2011). Our conclusion about the preferred nature of nonlinear waiting-time penalty contracts from the purchaser's standpoint is particularly important in view of the increasing complexity of emerging performance-based contract structures (UK Department of Health 2012). At present, standard NHS contracts explicitly penalize hospitals for violating the 18-week waiting-time target, though they do it through a detailed penalty-assessing mechanism, nonlinear in terms of the achieved waiting-time performance. Although our results on the performance of threshold contracts provide a promising starting point, more investigation is needed into the nature of nonlinear penalty contracts that can close the information-asymmetry-generated efficiency gap for an arbitrary patient mix.

Electronic Companion

An electronic companion to this paper is available as part of the online version at http://dx.doi.org/10.1287/msom.1120.0402.

Acknowledgments

The authors are grateful to many professionals at Addenbrooke's Hospital in Cambridge, United Kingdom, and the UK National Health System for their generous assistance in helping them 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 KJ (1963) Uncertainty and the welfare economics of medical care. *Amer. Econom. Rev.* 53(5):941–973.

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

Bolton P, Dewatripont M (2005) *Contract Theory* (MIT Press, Cambridge, MA).

Brun O, Garcia J (2000) Analytical solution of finite capacity $M/D/1$ queues. *J. Appl. Probab.* 37(4):1092–1098.

Cachon GP (2003) Supply chain coordination with contracts. Graves S, de Kok T, ed. *Handbooks in Operations Research and Management Science: Supply Chain Management* (Elsevier, Amsterdam), 229–339.

California Healthline (2011) California gears up to implement new rules on medical wait times. (January 5), http://www.californiahealthline.org/articles/2011/1/5/california-gears-up-to-implement-new-rules-on-medical-wait-times.aspx.#ixzz21IPWNcPn.

De Fraja G (2000) Contracts for health care and asymmetric information. *J. Health Econom.* 19(5):663–677.

Farrar S, Sussex J, Yi D, Sutton M, Chalkley M, Scott T, Ma A (2007) National evaluation of payment by results: Report to the Department of Health. Health Economics Research Unit, University of Aberdeen, Aberdeen, Scotland, UK.

Fuloria PC, Zenios SA (2001) Outcomes-adjusted reimbursement in a health-care delivery system. *Management Sci.* 47(6):735–751.

Goddard M, Mannion R, Smith P (2000) Enhancing performance in health care: A theoretical perspective on agency and the role of information. *Health Econom.* 9(2):95–107.

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

Gupta D, Wang L (2008) Revenue management for a primary-care clinic in the presence of patient choice. *Oper. Res.* 56(3):576–592.

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

Hasija S, Pinker EJ, Shumsky RA (2008) Call center outsourcing contracts under information asymmetry. *Management Sci.* 54(4):793–807.

Institute of Medicine (2007) *Rewarding Provider Performance: Aligning Incentives in Medicare* (National Academies Press, Washington, DC).

Integrated Healthcare Association (2011) Approved MY 2011 P4P measurement set. Accessed July 21, 2012, http://iha.org/pdfs_documents/p4p_california/ApprovedMY2011MeasureSet111610.pdf.

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

Kim S-H, Cohen MA, Netessine S (2007) Performance contracting in after-sales service supply chains. *Management Sci.* 53(12):1843–1858.

Lee DKK, Zenios SA (2012) An evidence-based incentive system for Medicare's End-Stage Renal Disease program. *Management Sci.* 58(6):1092–1105.

Levaggi L, Levaggi R (2010) Strategic costs and preferences revelation in the allocation of resources for health care. *Internat. J. Health Care Finance Econom.* 10(3):239–256.

Liu N, Ziya S, Kulkarni VG (2010) Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manufacturing and Service Oper. Management* 12(2):347–364.

Lu M, Donaldson C (2000) Performance-based contracts and provider efficiency: The state of the art. *Disease Management Health Outcomes* 7(3):127–137.

Mannion R, Marini G, Street A (2008) Implementing payment by results in the English NHS: Changing incentives and the role of information. *J. Health Organ. Management* 22(1):79–88.

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

Miraldo M, Siciliani L, Street A (2011) Price adjustment in the hospital sector. *J. Health Econom.* 30(1):112–125.

Mullen KJ, Frank RG, and Rosenthal MB (2010) Can you get what you pay for? Pay-for-performance and the quality of health care providers. *RAND J. Econom.* 41(1):64–91.

Patrick J, Puterman ML, Queyranne M (2008) Dynamic multipriority patient scheduling for a diagonostic resource. *Oper. Res.* 56(6):1057–1525.

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

Siciliani L (2007) Optimal contracts for health services in the presence of waiting times and asymmetric information. *B.E. J. Econom. Anal. Policy* 7(1):Article 40.

So KC, Tang CS (2000) Modeling the impact of an outcome-oriented reimbursement policy on clinic, patients, and pharmaceutical firms. *Management Sci.* 46(7):875–892.

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

Tian N, Zhang ZG (2006) *Vacation Queueing Models: Theory and Applications* (Springer Science + Business Media, New York).

UK Department of Health (2012). 2012/13 NHS standard contract for acute, ambulance, community and mental health and learning disability services (bilateral), Section B (the Services), p. B23. Accessed July 21, 2012, http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_131988.