Can Yardstick Competition Reduce Waiting Times?

Nicos Savva · Tolga Tezcan
London Business School, Regent's Park, London NW1 4SA, UK
nsavva@london.edu · ttezcan@london.edu
Ozlem Yildiz
Darden School of Business, University of Virginia, Charlottesville, Virginia 22903
yildizo@darden.virginia.edu

Yardstick competition is a regulatory scheme for local monopolists (e.g., hospitals), where the monopolist's reimbursement is linked to performance relative to other equivalent monopolists. This regulatory scheme is known to provide cost-reduction incentives and serves as the theoretical underpinning behind the hospital prospective reimbursement system used throughout the developed world. This paper uses a game-theoretic queueing model to investigate how yardstick competition performs in service systems (e.g., hospital emergency departments), where in addition to incentivizing cost reduction the regulator wants to incentivize waiting time reduction. We first show that the form of cost-based yardstick competition used in practice results in inefficiently long waiting times. We then demonstrate how yardstick competition can be appropriately modified to achieve the dual goal of cost and waiting-time reduction. In particular, we show that full efficiency (*first-best*) can be restored if the regulator makes the providers' reimbursement contingent on their service rates and is also able to charge a provider-specific "toll" to consumers. More importantly, if such a toll is not feasible, as may be the case in healthcare, we show that there exists an alternative and particularly simple yardstick-competition scheme, which depends on the average waiting time only, that can significantly improve system efficiency (*second-best*). This scheme is easier to implement as it does not require the regulator to have detailed knowledge of the queueing discipline. We conclude with a numerical investigation that provides insights on the practical implementation of yardstick competition for hospital Emergency Departments and also present a series of modelling extensions.

Key words: Yardstick competition, hospital regulation, emergency care, game theory, queueing theory.
History: March 22, 2018

1. Introduction

Services constitute a large part of the developed world economy and, in some cases, they operate as regulated monopolies. A case in point is the hospital industry, which in 2014 constituted 5.6% of the US economy, and is highly regulated by bodies such as the Centers for Medicare & Medicaid Services (CMS), which are responsible for approximately 45% of hospital reimbursements (CMS 2014).[1] Despite this, academic research on the regulation of service monopolies has not received

[1] In other developed world healthcare systems, such as the UK, hospitals are just as large a part of the national economy, and most of their reimbursement comes from a single government-funded payer that acts as the regulator.

as much attention as that of "production" monopolies, for example, defense systems, water, and energy (Laffont and Tirole 1993, Roques and Savva 2009).

This paper focuses on a specific regulatory scheme, yardstick competition, in which the regulator induces artificial competition between local monopolists by rewarding firms on the basis of their performance relative to each other. Yardstick competition has been shown to provide incentives for monopolists to minimize production costs (Shleifer 1985) and has been widely applied to the reimbursement of regulated utilities (e.g., electricity (Jamasb and Pollitt 2000) and tertiary care – through CMS's Diagnosis-Related Group (DRG) prospective reimbursement in the US and equivalent schemes in other developed economies (Fetter 1991)). This paper shows that yardstick competition fails to incentivize investment in wait-time reduction in service systems. In fact, this may be a contributing factor to the undesirably long waiting times observed in some service systems that are reimbursed through yardstick competition, e.g., hospital care (GAO 2009). To incentivize wait-time reduction through capacity investment or process re-engineering without compromising incentives for cost control, the yardstick competition scheme must be modified. This paper proposes one such modification.

What is yardstick competition? Historically, franchised monopolies had been subject to "cost-of-service" regulation, where the regulated firm's price is set equal to the (marginal) cost of production, along with a transfer payment, if needed, to ensure that the monopolist breaks even. The fee-for-service reimbursement model, which had been used by CMS in reimbursing US-based hospitals up to 1983, is one such example (Mayes 2007). Besides simplicity, the advantage of this regulatory scheme is that it avoids the higher monopoly price and the associated welfare loss, whilst providing incentives for the firm to continue production. The disadvantage is that it does not provide incentives to contain the cost of production.

A better alternative would be to dissociate the firm's price from the firm's cost of production, and instead set it equal to an exogenous benchmark. Setting this benchmark optimally, however, would require the regulator to know as much about the available cost-reduction technologies as the regulated firm. Yardstick competition gets around this difficulty by making use of the production cost of other equivalent monopolists to infer a firm's attainable costs, which will then serve as the exogenous benchmark. For example, the regulator could choose to reimburse the monopolist at the average production cost of all other monopolists. By doing so, the regulator forces the firms to engage in cost-reduction competition, akin to a tournament. Shleifer (1985) shows that, in the unique symmetric Nash equilibrium of this game, all firms invest in cost reduction optimally. That is, it achieves the same cost-reduction investment as that chosen by the regulator under full information. Since the price is set through observable and verifiable benchmarks, this scheme

can be implemented using accounting data without requiring symmetric information between the regulator and the regulated firms (Laffont and Tirole 1993).

Yardstick competition and waiting times. The theory outlined above abstracts away one important aspect of service provision: waiting times. Effectively, it assumes that production is instantaneous, or equivalently, that consumers' cost of waiting is negligible. This assumption may fit product-based monopolies, but it is less realistic for service settings where long waiting times are costly and, as in the case of hospital Emergency Departments (EDs), even dangerous. It is, therefore, important that, in addition to cost reduction, the regulation of service systems should also incentivize wait-time reduction.

To do this, we present a game theoretic queueing model of yardstick competition, where a regulator is responsible for multiple identical service providers that act as local monopolists. The regulator's objective is to maximize total welfare by dictating the price that customers are charged and any transfer payments made to the service providers. Service providers are assumed to be profit maximizers and, given the price and transfer payment set by the regulator, decide how much to invest in cost- and wait-time-reduction effort. Finally, customers are heterogeneous in their willingness to pay for the service, leading to an endogenous demand function that is decreasing in both the price as well as the waiting time they expect to incur. A crucial feature of our model, which explains why optimal regulation may be difficult, is that the providers' cost of investment in technology and the customer demand function are not known to the regulator. We note that extant models that ignore the cost of waiting (e.g., Shleifer 1985) are a special case of the model presented here.

Using this model, we show that, in contrast to product-based monopolies, applying the standard cost-based yardstick competition scheme to service monopolies results in inefficiently long waiting times, and may even result in inefficiently high costs due to two types of inefficiencies. First, customers will over-join the service system compared to first-best, a result first noted in Naor (1969). This is due to a negative externality: a customer's decision to join the system will lead to an increase in the expected waiting time of others. This increase does not feature in his personal calculation as to whether to join the system or not. Second, cost-based yardstick competition fails to generate incentives for the service providers to increase capacity.

To resolve these two inefficiencies, yardstick competition should be modified in two ways. First, the price that customers pay to access the service should be made higher than the costs of providing the service estimated through the yardstick competition benchmarks. This higher price constitutes a form of "toll" and is equal to the aggregate customer utility that a customer's joining decision displaces (Naor 1969). Second, the monopolists' reimbursement should have a component that

depends on their investment in wait-time reduction. More specifically, the monopolists' reimbursement should include a component that depends on the difference between her service rate and the average service rate of all other monopolists that serve as the benchmark. This modified yardstick competition achieves first-best outcomes on both cost and waiting-time reduction and, just like cost-based yardstick competition, can be implemented using accounting data on costs and waiting times, along with some information on the queueing discipline.

An essential feature of the yardstick competition scheme discussed above is that customers are charged a provider-specific fee for accessing the service. In many healthcare settings, this is not the case; care is either free at the point of access, or patients are charged a fixed insurance deductible. We show that, if customers are charged an exogenous fee (which could be zero), then there is an inefficiency due to the suboptimal customer joining behavior. Nevertheless, there is an alternative yardstick competition scheme that still achieves second-best, that is, there is no additional loss of welfare due to underinvestment in either cost or waiting-time reduction. Furthermore, this scheme is easier to implement – the regulator needs to observe the average customer waiting time and total provider costs, but does not need to know anything about the queueing discipline or how the costs are split between fixed and variable.

We also present a detailed numerical study that investigates the magnitude of the welfare loss associated with second-best outcomes compared to first-best, and the dependence of the equilibrium outcomes on the cost of waiting. One interesting finding is that, in contrast to equilibrium waiting times, total welfare is not very sensitive to the cost-of-waiting parameter. This suggests that the service provider may be able to use the cost-of-waiting parameter as a lever to shift the equilibrium outcome associated with the modified yardstick competition to one with lower waiting times albeit at higher costs, without significantly affecting total welfare.

Finally, in an electronic companion (EC), we present a series of extensions that explain how yardstick competition can be implemented in settings with demand-side competition between providers, multiple customer classes, time-varying arrival rates, more general cost structures, more general queueing disciplines including queueing networks. We also discuss how the regulator could adjust the yardstick competition scheme to account for heterogeneity between the local monopolists based on exogenous characteristics.

Applications. One immediate application of the modified yardstick competition model proposed by this paper is the regulation of hospitals and hospital EDs. First, prospective payments and cost-based yardstick competition have been widely adopted around the world as the primary mode of reimbursement for this industry.[2] Second, waiting time to receive emergency care is both costly

[2] The form of yardstick competition used in hospital reimbursement classifies acute patients into Diagnosis-Related

and, in many cases, excessively long.[3] This paper provides an explanation as to why there is such a systematic underinvestment in reduction of waiting times in emergency care. It also provides guidelines on how regulators could modify the reimbursement system already in use to provide incentives for waiting time reduction. More specifically, the second-best scheme that we propose is particularly useful, as it requires little additional information to implement (just average waiting time at the ED)[4] and achieves outcomes that cannot be improved upon without changing the way that patients are charged for accessing emergency care.

Beyond healthcare and hospital reimbursement, our work serves as an introduction to the notion of yardstick competition to the operations management community. More specifically, yardstick competition might be a useful tool for other service systems that operate as local monopolies (e.g., governmental agencies, such as the Department of Motor Vehicles and the US Social Security Administration offices; (former) quasi-state monopolies, such as the post office; airport or border security checkpoints), and for service firms to incentivize better performance for individual servers.

2. Literature Review

The observation that relative-performance evaluation is a useful tool for setting incentives has been made by Holmstrom (1982), Nalebuff and Stiglitz (1983), and Shleifer (1985) in related contexts: the first focuses on curbing free riding in teams, the second on optimal risk sharing, and the third on cost-cutting incentives for regulated firms. In fact, the term "yardstick competition" is used by Shleifer (1985) to describe this form of regulation. From a practical perspective, yardstick competition has been implemented in industries such as electricity production (Jamasb and Pollitt 2000) and water and sewage (Sawkins 1995).

Several extensions to the model of yardstick competition have been presented in the literature. For example, Laffont and Tirole (1993), pp. 84-86 augment the model of yardstick competition

Groups (DRGs) based on their diagnosis, existing complications and comorbidities, and patient-specific characteristics, for example, age (see Fetter (1991)). Patient episodes within a DRG require a similar bundle of services and goods to be diagnosed and treated. Hospitals are then reimbursed a fixed amount per patient that depends only on the patient's DRG. The amount is set to the average of the reported (and audited) cost of treating patients of the same DRG across all hospitals, adjusted for exogenous hospital characteristics such as local wages and training costs for teaching institutions. Since its introduction by CMS in the US in the 1980s, the system has been adopted by private insurance firms and healthcare payers throughout the developed world (Mayes 2007).

[3] In the US, patients who should have been seen in less than 1 minute waited for 28 minutes on average (GAO 2009). Similarly, ED wait times in England rose by one-third in November 2015 compared to November 2014 (Siddique 2016), and 10% of patients spent at least 8 hours in Canadian EDs (CIHI 2012). Furthermore, delays in the ED have been associated with a number of adverse outcomes, such as patient dissatisfaction, higher rates of medical errors, higher mortality rates, and more patients leaving without receiving treatment (Batt and Terwiesch 2015).

[4] In fact, regulators around the world have started collecting waiting-time data. For example, Monitor, the UK hospital regulator, collects data on ED waiting times and has placed an ad hoc target that at least 95% of patients have to be admitted or discharged within 4 hours of arriving at the ED with financial penalties for failure to comply (Siddique 2016). Similarly, CMS has started collecting data on ED waiting times, which are reported on the Hospital Compare website (CMS 2016b).

to regulate firms whose costs are imperfectly correlated. Sobel (1999) examines the case where transfers are costly to show that yardstick competition may discourage investment. This setting has been examined further in Lefouili (2015). Dada and White (1999) examine the financial risks associated with prospective payment systems that rely on yardstick competition in the context of hospital reimbursement. More relevant to our work are models that use prospective payments to incentivize improvements in additional dimensions of performance, such as quality. Examples include Ellis and McGuire (1986), Pope (1989), Ma (1994), and Tangerås (2009), where the general finding is that quality is better served by more complicated forms of yardstick competition. Our work adds to this literature because, unlike quality, waiting times are: i) endogenous to customer behavior and generate an externality on the customer side; ii) governed by well-understood non-linear dynamics that need to be accounted for; iii) often easier to quantify and less controversial to compare across providers than other quality measures (e.g., in-hospital mortality).

In addition to the literature on yardstick competition, this paper also contributes to the operations management (OM) literature that examines incentives and competition in queueing systems. Traditionally, queueing theory, which is well-surveyed in Kleinrock (1975), has been concerned with the mathematical description and optimization of queueing systems without considering customer behavior or agency issues on behalf of firm management. Early attempts to include such economic considerations are reviewed by Hassin and Haviv (2003) and more recent work in Hassin (2016). Our work brings together elements from: i) the literature on strategic customer behavior in monopolistic queueing systems, which was first studied in Naor (1969) for observable queues and extended to unobservable queues by Edelson and Hilderbrand (1975) and multiclass queues by Mendelson and Whang (1990) and Afeche (2013); and ii) the literature on queueing games where service providers compete based on price and congestion (e.g., Cachon and Harker (2002), Cachon and Zhang (2006), and Allon and Federgruen (2008)). Closest to our work are the studies of queueing games in the context of hospital/ED congestion (e.g., Lee and Cohen (1985)) and ambulance diversion (e.g., Deo and Gurvich (2011)). In contrast to this stream of literature, service providers in our setting do not compete directly. Instead, competition is induced by the regulator through the reimbursement mechanism. Furthermore, our work is complementary to the aforementioned papers on ambulance diversion, as it focuses on hospital reimbursement mechanisms that incentivize optimal investment in capacity, which, as a side effect, may make the need to divert ambulances less prevalent.

Our work is also related to the OM literature on performance-based incentives in services in general (Akan et al. 2011, Bakshi et al. 2015, Hasija et al. 2008, Kim et al. 2007, Kim et al. 2010, Ren and Zhou 2008) and in healthcare specifically (So and Tang 2000, Lee and Zenios 2012, Adida et al. 2016, Guo et al. 2016, Zorc et al. 2017, Andritsos and Aflaki 2015, Jiang et al. 2016). The last two papers also consider direct competition between providers (i.e., a queueing game) in the presence

of performance-based incentives. Our paper differs, as the performance-based incentives that we consider are not set exogenously by the regulator, but are the result of endogenous benchmarks. This is an important difference because it generates (indirect) competition between otherwise monopolistic providers and, as we show in this paper, may be easier to implement, as it places less onerous informational burden on the regulator.

3. Model Description

The model considers the interaction between three parties: the regulator, the service provider, and the customers. The regulator has $N \geq 2$ identical service providers under his jurisdiction and has the ability to set the price that customers are charged. The regulator may also decide to award an additional transfer payment to the service providers, which may depend on any observable and verifiable quantity. Customers observe how much they need to pay for service and decide whether to request service, which is provided on a first-come-first-served basis. As a result, customers may experience a costly wait, which we model explicitly using queueing theory. Finally, the service providers act as risk-neutral local monopolists. They observe the price and transfer payment set by the regulator and, given customer behavior, decide how much to invest in the cost- and wait-time-reduction effort. We present the details of the decisions and payoffs of each of the three parties and discuss how our model applies to the hospital emergency care setting in §§3.1-3.3.

3.1. Service Environment, Customer Utility and Equilibrium Arrival Rate

We assume that, within the catchment area of each service provider, there is a large population of customers who may experience a service need with an exogenous probability. The aggregate arrival rate of service needs may then be modeled as a Poisson process with rate Λ per unit time, even if customers are strategic, see Lariviere and Van Mieghem (2004). Each customer with a service need makes a decision whether to visit the service provider to receive service on a first-come-first-served basis, or use their outside option which, without loss of generality, we assume has a value of zero. In the case of emergency care, these assumptions reflect the case where patients have a single ED that they would consider visiting, either due to prohibitive transportation costs, informational frictions, or idiosyncratic preferences for a specific ED, for example, the closest (Brown et al. 2015). In this case, Λ would reflect the number of patients per unit time that exhibit a symptom (e.g., chest pain) for which they would consider visiting the ED. Since the patients exhibit the same symptom, they would all be classified in the same triage category; therefore, first-come-first-served is a reasonable assumption. If patients choose not to visit the ED, their outside options could be to use primary care or to not seek treatment. We present extensions to multiple customer classes, non-stationary arrivals, and demand-side competition between providers in the EC.

Each customer's utility from receiving service is comprised of three components. The first is the benefit from the service, r, which is net of any indirect costs associated with the service (e.g., net of transportation costs). We assume the value of service, r, to be heterogeneous across customers. The proportion of customers who value the service less than $x \geq 0$ is given by $\Theta(x)$. By definition, $\Theta(x)$ is non-negative and increasing. We also assume that its derivative, which we denote with θ, is strictly positive everywhere in $[0, \infty)$.[5] In the ED setting, the benefit, r, denotes the value that patients place on treatment and it is natural to assume that it is heterogeneous across patients due to the variability in the severity of patients' conditions, which is present even within the same triage category. The second component of customers' utility is the price of the service, p. In the ED setting, this may reflect the co-payment for an ED visit, which may well be zero. The third component is the cost of waiting to receive service (similar to Dai et al. (2016) and Wang et al. (2010)), which we assume to be t per unit time. In general, it reflects opportunity cost, and in the ED setting in particular, it may also reflect the monetary value of the anxiety, pain, and inconvenience that patients might experience until they are diagnosed and/or treated. We assume heterogeneity in this cost to be less pronounced than that in the benefit from receiving the service and, for tractability purposes, we model this as homogeneous across customers. Throughout we model customers' cost of waiting as a linear function of the waiting time, although, with minor modifications, our results hold for any convex increasing function of the waiting time as well.

More formally, the utility that each customer expects to receive from seeking service is given by $r - tW(\lambda, \mu) - p$, where $W(\lambda, \mu)$ denotes the expected waiting time, given the rate of customers arriving to the service provider, λ, and the actions of the service provider that result in increasing throughput, which are summarized by the variable μ (see also §3.2). Throughout, we assume that $W(\lambda, \mu)$ is increasing in λ and decreasing in μ, and that for any $\lambda \in (0, \mu)$

$$W(\lambda, \mu) > W(0, \mu) \text{ and } \lim_{\lambda \to \mu} W(\lambda, \mu) = \infty. \tag{1}$$

These reasonable assumptions imply that some delay is inevitable, and it is not possible to run the system close to 100% utilization without excessive delays. These assumptions clearly hold for the $M/M/1$ queue and for any multiple-server queue with random service and/or interarrival times.

Any customer with positive utility will seek service, and the equilibrium arrival rate, $\lambda(p, \mu)$, is given by the unique solution of the equation

$$\lambda(p, \mu) = \Lambda \bar{\Theta} \left(p + tW(\lambda(p, \mu), \mu) \right), \tag{2}$$

[5] To avoid subtle technical questions and to facilitate game-theoretic analysis, we assume that all functions defined are twice differentiable.

where $\bar{\Theta}(r) := 1 - \Theta(r)$ and $\lambda(p,\mu) < \mu$. If, for example, the service is provided in an $M/M/1$ queue, this equation reduces to

$$\lambda(p,\mu) = \Lambda\bar{\Theta}\left(p + \frac{t}{\mu - \lambda(p,\mu)}\right). \tag{3}$$

We note that the formulation above assumes that customers do not observe the actual waiting time when they make the decision to seek service. This is consistent with many practical settings, including EDs where patients have little visibility of actual waiting times before they visit the ED (see Chapter 3 of Hassin and Haviv (2003) for an excellent review of the literature on unobservable queues and its applications). Nevertheless, customers are assumed to have accurate beliefs about expected waiting times, which they may have formed through repeated interactions with the service provider, word of mouth, or online tools that publish average ED waiting times (CMS 2016b).

3.2. Service Provider's Profit and Actions

We next discuss the profit maximization problem of one service provider of the N identical service providers. For simplicity, and in order to generate results that are comparable with extant literature, we present a single-period model where the reimbursement mechanism, which consists of a customer price, p, and transfer payment, T, is set by the regulator at the beginning of the period. The duration of this period is much longer than the average patient-interarrival time. Given the reimbursement mechanism and the customers' equilibrium arrival rate, $\lambda(p,\mu)$ given in (2), the service provider's profit per unit time (throughout the time period) is given by

$$\Pi(c,\mu|p,T) = (p-c)\lambda(p,\mu) - R(c,\mu) + T, \tag{4}$$

where c is the cost of providing service per customer and μ represents the level of effort that the service provider chooses to exert in order to reduce waiting time. The cost function, $R(c,\mu)$, denotes the cost of all activities undertaken by the service provider to reduce the cost of providing service to the level, c, and the cost of effort, μ, associated with reducing the waiting time. We assume that cost, $R(c,\mu)$, is a fixed cost, at least in the short-run, and is decreasing in cost per customer, c, increasing in effort, μ, and it is jointly convex.

In the case of the ED, the single period of time may represent a year within which the regulator has committed not to make any further adjustments to the regulatory environment. The cost per customer, c, represents the overall cost of treating a patient with a specific condition, and $R(c,\mu)$ denotes the cost (per unit time) of any interventions or process re-engineering that may yield a more cost-efficient process or increase in throughput. For instance, purchasing capital intensive new equipment that allows for more precise and faster patient treatment, employing more and better-qualified staff, and/or re-engineering processes (e.g., having patients triaged and diagnosed by more

experienced healthcare providers (Saghafian et al. 2014), can reduce the cost of treating patients and simultaneously increase the throughput (see Saghafian et al. (2015) for more on throughput improvements in EDs.) We extend this model to more general cost structures in the EC. If the service is provided in an $M/M/1$ queue, the variable μ can be interpreted as the service rate (or service capacity) of the system and, for this reason, we will refer to μ as effort or capacity interchangeably.

At the beginning of the period, the provider chooses the optimal cost, c, and waiting time reduction effort, μ, by solving the profit maximization problem

$$\max_{0<c\leq c_o, 0<\mu_o\leq\mu} \Pi(c,\mu|p,T). \tag{5}$$

We assume that the service provider must choose the cost per patient, c, from the interval $[0, c_o]$ and the capacity level, μ, from the interval $[\mu_o, \infty)$. The objective of profit maximization is not inconsistent with hospital care, see for example the discussion in Andritsos and Aflaki (2015). Naturally, to solve the problem specified above, the provider must know the cost function, $R(c,\mu)$, and be able to estimate the patient demand, $\lambda(p,\mu)$ (which requires knowing the queueing dynamics, the distribution of customers' valuation, and the size of catchment area). The limits, c_o, which can be arbitrarily large, and μ_o, which can be arbitrarily small, can be thought of as the (exogenous) default cost and capacity decisions of the provider.

Finally, we note that our model includes N non-competing providers who we assume are identical in terms of their profit functions. We extend our analysis to heterogeneous providers and to providers who compete on waiting times in the EC.

3.3. Regulator's Welfare

The regulator has the authority to dictate the price, p, paid by customers, the transfer, T, received by the providers, and may also choose to dictate the cost, c, and capacity, μ, set by the service providers. These are chosen at the beginning of the time horizon in order to maximize total welfare, which comprises the accumulated customer surplus and the profits of each of the N service providers. The expression for the total welfare rate associated with one such service provider is given by

$$S(p,c,\mu) = \Lambda \int_{p+tW(\lambda,\mu)}^{\infty} (x - p - tW(\lambda,\mu))\, d\Theta(x) + (p-c)\lambda(p,\mu) - R(c,\mu), \tag{6}$$

where $\lambda(p,\mu)$ is given in (2). The first term in the expression above is the total consumer surplus per unit of time. The second and third terms together constitute the profit of the service provider, net of the transfer payment. We note that the transfer payment, T, does not appear in the welfare function as it is a payment within the system. Nevertheless, the transfer payment may be necessary

to ensure that the service provider breaks even, that is, $\Pi(c,\mu|p,T) \geq 0$, and would therefore continue to provide service. Under the $M/M/1$ assumption, the social welfare rate can be written as

$$S(p,c,\mu) = \Lambda \int_{p+\frac{t}{\mu-\lambda(p,\mu)}}^{\infty} \left(x - p - \frac{t}{\mu-\lambda(p,\mu)}\right) d\Theta(x) + (p-c)\lambda(p,\mu) - R(c,\mu), \tag{7}$$

where $\lambda(p,\mu)$ is given by (3).

In the ED setting, we assume that the role of the regulator is fulfilled by the main payer (e.g., CMS in the US or the national payer in other more centralized systems) whose objective is to maximize the sum of patient utility and hospital profits. Similar objectives have been used extensively in healthcare economics and operations management literature, e.g., Andritsos and Tang (2015), Adida et al. (2016).

3.4. First-best Benchmark

We start the analysis by finding the welfare maximizing price, p, transfer payment, T, cost per customer, c, and capacity, μ, assuming that the regulator has full information about all model parameters, including the cost function, $R(c,\mu)$, of the service provider and the equilibrium arrival rate, $\lambda(p,\mu)$, of the customers. In this centralized setting the regulator solves

$$\max_{p\geq 0, c_o \geq c > 0, \mu \geq \mu_o > 0, T} S(p,c,\mu) \tag{8}$$

$$\text{s.t. } \Pi(c,\mu|p,T) \geq 0. \tag{9}$$

We highlight that, given any level of price, p, cost per customer, c, and capacity, μ, any transfer payment, T, above a threshold would satisfy the provider's break-even constraint in (9). Here, we implicitly assume that the regulator prefers reimbursing the provider as little as possible while ensuring that (9) holds (see also Sobel (1999)). We also note that, due to the complicated (and endogenous) queueing dynamics, the welfare function might not always be concave. As usual in the literature of queueing games, we assume that first-order conditions (FOCs) are necessary and sufficient for determining the unique solution to the regulator's welfare maximization problem. We present sufficient conditions for this to be the case in the EC.

Proposition 1. *The unique welfare-maximizing (first-best) price, p^*, cost per customer, c^*, capacity, μ^*, and transfer payment, T^*, are given by*

$$\frac{\partial}{\partial c} R(c^*,\mu^*) = -\lambda^*, \tag{10}$$

$$\frac{\partial}{\partial \mu} R(c^*,\mu^*) = -t\lambda^* \frac{\partial}{\partial \mu} W(\lambda^*,\mu^*), \tag{11}$$

$$p^* = c^* + t\lambda^* \frac{\partial}{\partial \lambda} W(\lambda^*,\mu^*), \tag{12}$$

$$T^* = R(c^*, \mu^*) - t\lambda^{*2}\frac{\partial}{\partial\lambda}W(\lambda^*, \mu^*), \tag{13}$$

where $\lambda^* = \lambda(p^*, \mu^*)$ *is given by* (2). *In the* $M/M/1$ *case,* $-t\lambda^*\frac{\partial}{\partial\mu}W(\lambda^*, \mu^*) = t\lambda^*\frac{\partial}{\partial\lambda}W(\lambda^*, \mu^*) = \frac{t\lambda}{(\mu-\lambda)^2}$.

Proof of all results presented in the Appendix.

The solution to the regulator's problem makes intuitive sense. First, the transfer payment of (13) is such that the service provider breaks even. Second, the first-best service cost, c^*, given by (10) is set so that the marginal benefit from a reduction in the treatment cost across all customers seeking service, $\lambda\Delta c$, is equal to the marginal cost of cost reduction, $\frac{\partial R}{\partial c}\Delta c$. Third, the first-best service capacity, μ^*, given by (11), is set so that the marginal cost of increasing capacity, $\frac{\partial R}{\partial \mu}\Delta\mu$, is equal to the reduction in waiting time associated with the increase in capacity experienced by all customers who choose to seek service, $-t\lambda\frac{\partial W}{\partial\mu}\Delta\mu$. Fourth, the first-best price, p^*, given by (12), makes the customers who choose to seek service bear the cost of providing the service, c, plus an additional "toll" which is equal to the marginal externality cost incurred by their fellow customers due to the increase in waiting time, $t\lambda\frac{\partial W}{\partial\lambda}$.

We note that, if the cost of waiting is zero ($t = 0$), our results coincide with those of earlier models where waiting is assumed not to be costly, for example, Shleifer (1985). (In this case one can ignore condition (11) as capacity has no impact on welfare.) Comparing the setting with positive waiting costs to a setting where waiting costs are zero, we note one important difference: in the former, customers are charged more than the cost of providing service, that is, $p > c$. This result is similar to Naor (1969). The additional charge reflects the endogenous nature of waiting costs; that is, by joining the service provider, consumers make it more expensive for anyone else to join. They, therefore, have to be charged an additional "toll" to incentivize optimal joining behavior. As a consequence of this toll, the break-even transfer payment required is less than the investment cost, $R(c^*, \mu^*)$. Throughout this paper, we focus on the more interesting case where $\mu^* > \mu_o$ and $c^* < c_o$.

4. Regulatory Schemes

To implement the welfare maximizing capacity, μ^*, cost per customer, c^*, and price, p^*, the regulator needs to have perfect knowledge of the cost function $R(c, \mu)$, and the customer equilibrium arrival rate $\lambda(p, \mu)$. In practice, regulated firms often have privileged information vis-à-vis the regulator (Armstrong and Porter 2007). In the hospital setting specifically, the number of conditions treated and the pace of technological change associated with treatments would make it difficult for the regulator to maintain an accurate understanding of the costs. Similarly, the regulator may be less able to estimate the distribution of customers' benefit from service $\Theta(.)$, a critical input into

the calculation of the equilibrium arrival rate $\lambda(p,\mu)$, vis-à-vis the service provider who regularly interacts with the customers.

In contrast, the regulator may be able to observe audited accounting data on the cost of treatment, c, investment cost, R, and the average number of customers served per unit time, λ, along with the average waiting time, W, after the service provider chooses the capacity and marginal cost levels. For example, CMS already collects and audits the first three figures for all hospitals in the US and has recently started collecting the last. Similarly, in addition to costs, ED waiting times are also monitored by the UK hospital regulator. Motivated by this observation, we will present four regulatory regimes that do not assume knowledge of the cost function, $R(c,\mu)$, or the customer equilibrium arrival rate function, $\lambda(p,\mu)$. The first two, cost-of-service regulation and cost-based yardstick competition, have been implemented in practice, but their performance in a service setting, where waiting times are costly, has not been assessed before. The third and fourth regulatory regimes, which modify cost-based yardstick competition, are, to the best of our knowledge, new. For each of the regulatory schemes that we introduce, we need to make specific assumptions about the providers' profit function to ensure sufficiency of FOCs. As in the previous section, we present these sufficient conditions in the EC.

4.1. Cost-of-service Regulation

Under cost-of-service regulation, the service provider is free to decide on the capacity, μ, cost per customer, c. The reimbursement (in the form of price, p, and transfer payment, T), chosen by the regulator, is designed to cover the total cost of the service provider while avoiding distortions associated with the monopolist price that the service provider would naturally be inclined to impose. More specifically, under this scheme, which is similar to the way that hospitals were reimbursed by CMS until 1983 (Mayes 2007), the regulator would audit the service provider to determine the costs c and $R(c,\mu)$ and would then impose a customer price $p=c$ and transfer payment $T=R(c,\mu)$. Clearly, this scheme cannot induce socially optimal investment because the service provider makes zero profit regardless of the capacity and cost-reduction effort that it makes and, therefore, has no incentives to invest in either. This result is also noted in Shleifer (1985).

4.2. Cost-based Yardstick Competition

The reason why "cost-of-service" regulation fails to provide cost-reduction incentives is the dependence of the firm's reimbursement on its own chosen cost structure. An alternative regulatory scheme, which eliminates this dependence, has been proposed by Shleifer (1985). In his setting, customers do not experience costly waiting times, and the proposed regulatory regime achieves socially optimal levels of cost-reduction effort without relying on the regulator knowing the cost function of the firm, $R(c,\mu)$. Since this scheme is similar to the DRG payment system implemented

by CMS in hospital reimbursement, it is important to investigate if it can achieve socially desired outcomes in the case where consumers' delays are costly. Before we discuss this, we first explain the scheme proposed by Shleifer (1985). To do so requires defining some additional notation.

We follow the same notation as before but add a subscript i, which stands for the service provider index, $i=1,...,N$. For each service provider i, $i=1,...,N$, define

$$\bar{c}_i = \frac{1}{N-1}\sum_{j\neq i} c_j, \quad \bar{R}_i = \frac{1}{N-1}\sum_{j\neq i} R(c_j,\mu_j). \tag{14}$$

Under the scheme proposed by Shleifer (1985), the regulator sets customer price and transfer payment for provider i to be $p_i=\bar{c}_i$, and $T_i=\bar{R}_i$, respectively. Based on the price, p_i, and the capacity choice of service provider i, which determines its expected wait time, customers seek service at provider i with the rate given in (2). Since the price and transfer payment of each provider depend on the actions of all other providers, they are forced to engage in a simultaneous-move game with complete information, where each provider chooses the capacity, μ_i, and cost per customer, c_i, to maximize the payoff function given in (4). We present the equilibrium of this game below.

PROPOSITION 2 **(Shleifer 1985)**. *In the absence of costly waiting time ($t=0$), if the regulator sets provider's i's price and transfer payments by* (14)*, the unique Nash equilibrium is for each provider i to choose $c_i=c^*$, $i=1,...,N$, $N\geq 2$. Also, all providers make zero profit in equilibrium.*

By implementing the regulatory scheme described above, the regulator forces the providers to engage in indirect competition to reduce costs. This is achieved by first, decoupling the reimbursement rate of each provider from the cost chosen by the provider, and second, setting the reimbursement rate equal to an exogenous industry-wide benchmark cost level. In the absence of costly waiting time ($t=0$), the unique symmetric Nash equilibrium of this tournament-style competition generates first-best outcomes, that is, it achieves the same cost-reduction investment as that chosen by the regulator under full information derived in §3.4. However, this scheme can be implemented using cost-accounting data, and therefore, does not require symmetric information between the regulator and the regulated service providers. Furthermore, under this scheme all service providers achieve zero profits in equilibrium and, as a result, there is no reason for the regulator to want to renegotiate any payments after investments have been made, thus alleviating any concerns for hold-up problems (Sobel 1999).

Since this scheme depends only on the cost of providing the service (and not the level of capacity investment), we refer to this scheme as cost-based yardstick competition. We next investigate the performance of this scheme in the setting where customers are sensitive to delays.

PROPOSITION 3. *If customers experience costly waiting time ($t > 0$) under the cost-based yardstick competition of Proposition 2, the providers' installed capacity is the minimum capacity level $\mu_o < \mu^*$ in all potential symmetric equilibria. If, in addition, $\frac{\partial^2 R(c,\mu)}{\partial \mu \partial c} \geq 0$ for all $0 < c \leq c_o$ and $\mu \geq \mu_o$, then this reimbursement scheme results in a unique symmetric equilibrium where providers choose a higher cost compared to the first-best, c^*.*

This proposition shows that, in the presence of costly waiting time, cost-based yardstick competition results in underinvestment in capacity. This result, which holds irrespective of the detailed queueing discipline employed by the service provider, arises because, in equilibrium, the service provider has no incentive to increase capacity. Adding capacity investment is costly; however, the service provider does not receive any direct benefit from the associated reduction in waiting times (her payment is not linked to capacity or waiting time) or indirect benefit (although the reduction in waiting time will increase the equilibrium arrival rate, this will not increase the service provider's profit, as the marginal profit for each additional customer is, in equilibrium, zero). Furthermore, this scheme suffers from an additional source of inefficiency – given the capacity and marginal cost-reduction investment, more customers choose to seek service than is socially optimal (for *that* price and capacity level) in a similar spirit to Naor (1969). This can be seen by noting that the first-best price (see Proposition 1) is greater than the marginal cost (see Proposition 3). In addition to underinvestment in capacity, Proposition 3 shows that, if marginal cost reduction gets cheaper for providers with higher capacities (i.e., $\frac{\partial^2 R(c,\mu)}{\partial \mu \partial c} \geq 0$), cost-based yardstick competition leads to underinvestment in cost reduction as well (i.e., the cost per customer is greater than first best).

Clearly, cost-based yardstick competition falls short of achieving socially desired outcomes in a setting with capacity-constrained providers and delay-sensitive customers. The systematic underinvestment in capacity that arises as an equilibrium result from this scheme may be a contributing factor in the long waiting times observed in accessing emergency healthcare throughout the developed world (GAO 2009). For the rest of this paper, we investigate whether this shortcoming of cost-based yardstick competition can be improved by implementing alternative regulatory schemes.

4.3. Cost- and Capacity-based Yardstick Competition: First-best

We next propose a regulatory scheme that incentivizes service providers to take the first-best actions established in Proposition 1 when waiting time is costly, that is, $t > 0$. Let λ_i and μ_i respectively denote the arrival rate and capacity of service provider i and define

$$\bar{\lambda}_i = \frac{1}{N-1} \sum_{j \neq i} \lambda_j \quad \text{and } \bar{\mu}_i = \frac{1}{N-1} \sum_{j \neq i} \mu_j, \tag{15}$$

for $i = 1, ..., N$. Consider the following payment scheme: each customer seeking service from provider i is charged a price p_i, where

$$p_i = c_i + t\lambda_i \frac{\partial}{\partial \lambda} W(\lambda_i, \mu_i). \tag{16}$$

In addition, the regulator sets the transfer payment T_i to provider i as

$$T_i = (\bar{c}_i - c_i)\bar{\lambda}_i + t\bar{\lambda}_i \frac{\partial}{\partial \mu} W(\bar{\lambda}_i, \bar{\mu}_i)(\bar{\mu}_i - \mu_i) + \bar{R}_i - t\lambda_i^2 \frac{\partial}{\partial \lambda} W(\lambda_i, \mu_i). \tag{17}$$

Under this payment scheme, service provider i's objective function can be written as

$$\Pi(c_i, \mu_i | p_i, T_i) = (\bar{c}_i - c_i)\bar{\lambda}_i + t\bar{\lambda}_i \frac{\partial}{\partial \mu} W(\bar{\lambda}_i, \bar{\mu}_i)(\bar{\mu}_i - \mu_i) - R(c_i, \mu_i) + \bar{R}_i. \tag{18}$$

As in the case of cost-based yardstick competition, this payment scheme induces a simultaneous-move game between the providers. We investigate the equilibrium outcome of this game with the theorem below.

THEOREM 1. *If the regulator sets service provider i's price equal to p_i given in* (16) *and transfer payment equal to T_i given in* (17)*, then the unique symmetric Nash equilibrium is for each provider i to pick $c_i = c^*$ and $\mu_i = \mu^*$, for $i = 1, ..., N$. Also, all providers make zero profit in equilibrium.*

The regulatory scheme proposed in this section consists of a per-customer price, p_i, and a transfer payment, T_i, similar to cost-based yardstick competition. In addition to costs, each now also depends on the capacity decision, μ_i, directly. The price, p_i, which is equal to the cost of providing the service, c_i, plus the expected waiting-cost externality ($t\lambda_i \frac{\partial}{\partial \lambda} W(\lambda_i, \mu_i)$), serves the purpose of regulating the customers' joining behavior and, in equilibrium, is equal to the first-best price, p^*. Without it, customers would over-join compared to the socially optimal arrival levels as explained in §3.4. The transfer payment, T_i, coupled with the fee paid by each customer, serves to align the providers' incentives with the regulators' and, at the same time, ensures that the providers break even. The first term of the transfer payment ($(\bar{c}_i - c_i)\bar{\lambda}_i$), which is decreasing in the difference between the costs of provider i and the industry average, puts pressure on each provider to reduce costs to first-best levels. The second term ($t\bar{\lambda}_i \frac{\partial}{\partial \mu} W(\bar{\lambda}_i, \bar{\mu}_i)(\bar{\mu}_i - \mu_i)$), which is increasing in the difference between the service rate of provider i and the rest of industry that serves as a benchmark ($\mu_i - \bar{\mu}_i$), provides the right incentives for each provider to increase capacity to first-best levels. The final two terms serve to ensure that the providers break even in equilibrium, thus alleviating any concerns that the contracts may be renegotiated. Furthermore, in equilibrium, all but the last two terms in the transfer payment would simplify to zero; thus, the actual equilibrium payment would simplify to $\bar{R}_i - t\lambda_i^2 \frac{\partial}{\partial \lambda} W(\lambda_i, \mu_i)$, or, in the case of $M/M/1$ queueing discipline to $\bar{R}_i - \frac{t\lambda_i^2}{(\mu_i - \lambda_i)^2}$.

We note that the scheme proposed in this section achieves first-best without requiring the regulator to have symmetric information about either the cost function, $R(c, \mu)$, or the customer equilibrium arrival rate function, $\lambda(p, \mu)$. Nevertheless, we think that it would be difficult to implement in practice, especially in the case of hospital ED regulation. First, the mechanism proposed above requires customers to be charged a provider- and condition-specific fee to achieve socially optimal

arrivals. This might be possible in certain industries; however, in most healthcare delivery systems, patients do not bear the cost of treatment directly. For example, in the UK healthcare is funded through taxes and is free of charge to all residents. Although in other healthcare systems, such as the US, patients may be required to pay a fee when they receive treatment (e.g., in the form of co-payments), this fee is not tied to the performance of the provider and does not depend on the patient's condition. Second, in order for the regulator to implement the yardstick competition mechanism proposed in this section, it is necessary to have some information about the queueing discipline at the providers' sides. This is needed in order to estimate the service capacity, μ, installed by each provider and the waiting time function and its derivatives with respect to the arrival rate, λ, and capacity, μ. We suspect that in the highly complex hospital ED setting the queueing discipline would be hard to observe for the regulator. Motivated by the first practical difficulty above, in the next section we provide an alternative scheme which does not charge the customers a provider- and condition-specific fee. Fortunately, as we show in the next section, this scheme also alleviates the second concern as well.

We conclude this section by noting that we cannot rule out the existence of asymmetric equilibria. This is a common problem in such settings, see for example, Shleifer (1985). Nevertheless, in § EC.4 we present a more complex regulatory scheme for which we can also rule out the existence of any asymmetric equilibria.

4.4. Free-at-the-point-of-care Yardstick Competition: Second-best

To address the concern that it is often not feasible to charge customers directly, in this section, we propose an alternative payment scheme that guarantees that the chosen actions of the service providers will maximize welfare. For the rest of this section, we assume that the regulator charges a fixed price, which we fix to $p=0$ and drop from the notation, for example, we set $\lambda(\mu)=\lambda(0,\mu)$, with a slight abuse of notation. The analysis of this section would be almost identical if customers were charged a fixed fee, as in the case of patient co-payments for visiting EDs.

First, consider the objective function of the regulator $S(c,\mu)$ defined as in (6) with $p=0$. The optimal solution to this problem which we label as "second-best" solution and denote with μ_o^* and c_o^*, is given by the following proposition.

Proposition 4. *The unique welfare-maximizing (second-best) capacity μ_o^*, cost per customer c_o^*, and transfer payment T_o^*, when price $p=0$ are given by*

$$\frac{\partial}{\partial c}R(c_o^*,\mu_o^*)=-\lambda(\mu_o^*), \tag{19}$$

$$\frac{\partial}{\partial \mu}R(c_o^*,\mu_o^*)=-t\lambda(\mu_o^*)\frac{d}{d\mu}W(\lambda(\mu_o^*),\mu_o^*)-c\lambda'(\mu_o^*), \tag{20}$$

$$T_o^*=R(c_o^*,\mu_o^*). \tag{21}$$

In the $M/M/1$ case, $\frac{d}{d\mu}W(\lambda(\mu),\mu)=\frac{\lambda'(\mu)-1}{(\mu-\lambda(\mu))^2}$.

We note that, in the absence of a provider-specific fee (e.g., $p=0$), consumer behavior is going to be inefficient – some customers with a sufficiently low valuation who would have chosen not to visit the provider under first-best price, p^*, will now find it optimal to seek service. In fact, when $p=0$, the only reason that not everyone seeks service is congestion – some potential customers find the cost of the (equilibrium) expected waiting time to outweigh the benefit from receiving service.

For each service provider i, we define the average waiting time of all other service providers as

$$\bar{W}_i = \frac{1}{N-1}\sum_{j\neq i} W(\lambda(\mu_j),\mu_j),\ i=1,\ldots,N. \tag{22}$$

For notational simplicity, we set $W_i := W(\lambda(\mu_i),\mu_i)$. Consider the payment scheme where the regulator pays provider i a transfer payment equal to

$$T_i = t(\bar{W}_i - W_i)\bar{\lambda}_i + \bar{R}_i + \bar{c}_i\bar{\lambda}_i. \tag{23}$$

Under this payment scheme, service provider i's objective function is given by

$$\Pi(c_i,\mu_i|T_i) = -c\lambda(\mu_i) + t(\bar{W}_i - W_i)\bar{\lambda}_i - R(c_i,\mu_i) + \bar{R}_i + \bar{c}_i\bar{\lambda}_i. \tag{24}$$

The payment scheme defined above forces the service providers to engage in a simultaneous-move game whose equilibrium outcome we present below.

THEOREM 2. *If the regulator makes transfer payment T_i defined as in* (23) *to provider i, for $i=1,\ldots,N$ and customers are not charged directly, the unique symmetric Nash equilibrium is for each provider i to pick $\mu_i=\mu_o^*$ and $c_i=c_o^*$ for $i=1,\ldots,N$. Also, all providers make zero profit in equilibrium.*

The implication of Theorem 2 is that, in the absence of a direct customer fee, yardstick competition is still useful. Although it cannot restore first-best outcomes (as there is no way to counter the inefficient joining behavior of customers), by implementing the scheme proposed above, the regulator can achieve the second-best outcome, even though he has no information about the cost structure of the service providers, $R(c,\mu)$, or the customer equilibrium arrival rate function, $\lambda(p,\mu)$. The incentive to invest optimally (in the second-best sense) in capacity, μ, comes from the transfer payment, which is an increasing function in the difference between the industry benchmark waiting time, $\bar{W}_i$, and that chosen by the provider, W_i. This creates the tournament-style incentives that lead to a unique symmetric equilibrium where all providers invest optimally in capacity. Similarly, each provider has an incentive to invest optimally in cost reduction (again, in a second-best sense) as the payment scheme described above, which pays the provider a fee that is independent of their own actions, makes the provider the residual claimant – the additional value generated by lower costs is fully appropriated by the provider.

Furthermore, the scheme proposed in this section is simpler than that proposed in §4.3, where customers are charged a direct fee, for three reasons. First, it requires no information on the service rate, μ, or the queueing discipline and the associated waiting time function, $W(\lambda, \mu)$, and its derivatives. Instead, the only additional requirement compared to the simpler cost-based yardstick competition of §4.2 is that the regulator also monitors the average wait time for each provider. Second, the equilibrium transfer payment is equal to the total cost incurred by the service provider $(R_i + c_i\lambda_i)$, which, in contrast to the transfer payment of the first-best scheme of §4.3, is always non-negative. Third, it does not require that the regulator is able to separately observe the cost of providing service, c, and the investment costs, R. Instead, it suffices to observe the total cost incurred by each provider $R_i + c_i\lambda_i$ (see also Meran and Von Hirschhausen (2009)), which is a simpler task in many cases where variable and fixed costs are not easy to delineate (such as hospital care, see for example, Freeman et al. (2016)). For these reasons, we expect this scheme to be easier to implement in practice than the first-best scheme of §4.3.

However, we note that, despite its simplicity, this scheme requires that the regulator knows the patients' cost of waiting, t, which may not always be the case. We investigate this dependence numerically in §5. Furthermore, the simplicity of the second-best yardstick competition comes at a cost of efficiency. The loss of efficiency, which we also investigate numerically in §5, is due to the suboptimal customer joining behavior, which this regulatory scheme does nothing to curtail. In that sense, this regulatory scheme treats waiting times as any other exogenous quality measure that the regulator might care about (e.g., hospital readmission rates (see Zhang et al. (2016)) or adherence to best-practice protocols (see Gaynor (2004) for a literature review and background) and augments the standard yardstick competition of §4.2 in order to provide sufficient incentives to invest optimally in improving this quality measure. Therefore, and perhaps not surprisingly, the scheme proposed in this section has some similarities to a scheme already in use by CMS to provide quality improvement incentives in dimensions other than costs (e.g., Hospital Value-Based Purchasing program (CMS 2016a) or the Hospital Readmission Reduction Program (Zhang et al. 2016)).

We conclude this section by noting that Theorem 2 does not rule out the existence of asymmetric equilibria. Nevertheless, we are able to show in the §EC.5 that, when there are only two providers (i.e., $N = 2$), the symmetric equilibrium is indeed unique. Using this observation, we can then propose an alternative mechanism that does result in a unique equilibrium which leads to second-best outcomes. In this mechanism, providers are divided into two disjoint sets, and the average performance of one set is used to set a yardstick for the other and vice versa.

5. Numerical Investigation

In §4, we have argued that the second-best yardstick competition would be easier to implement in healthcare settings such as the regulation of EDs, as: a) customers are typically not charged a provider-specific fee for accessing care; and b) it places a lower informational burden on the provider. These advantages come at the cost of not achieving first-best level of investment in either wait-time or cost reduction. In this section, we numerically investigate the efficiency loss associated with this second-best outcome. In addition, we also investigate the equilibrium cost of second-best regulation and the impact of error in the estimation of one critical model parameter, the cost of waiting, t.

5.1. Model Parameters

The queueing model presented in this section is clearly a stylized representation of reality. Nevertheless, we have chosen a range of parameter values that match, as far as possible, the ED setting.

- We set the cost of waiting, t, to \$30 per hour which is approximately the 75th percentile of US hourly wages (Bureau of Labor Statistics 2016). We vary this from \$10 to \$100 per hour, a range which contains more than 80% of the population's hourly wages.

- We assume that the distribution of patients' benefit from treatment follows the exponential distribution $\Theta(x) = 1 - e^{-\alpha x}$, where x is benefit from service (in dollars) and $\alpha > 0$. This implies that the price elasticity of demand is $-\alpha x$. To estimate the elasticity parameter α we use the fact that (i) at the average cost, US healthcare price elasticity is estimated to be -0.17 (Ringel et al. 2002); and (ii) the average cost is approximately equal to \$200 – this is the sum of the average co-payment for an ED visit, estimated to be \$140 (CEB 2016), and the average cost of waiting (which is given by multiplying the average ED waiting time of two hours, as reported in Batt and Terwiesch (2015), with the cost of waiting of \$30 per hour). This generates a base estimate of $\alpha = 8.5 \times 10^{-4}\$^{-1}$. We run a sensitivity analysis for α ranging from $5 \times 10^{-4}\$^{-1}$ to $20 \times 10^{-4}\$^{-1}$, which corresponds to price elasticity ranging from -0.10 to -0.40.

- To estimate the size of the total potential demand Λ (i.e., the demand if waiting times and price were both zero), we start from the observation that, at current waiting times, average realized ED demand in the US in 2011 was 44.5 visits per 100 persons per year (CMS 2011). We also assume that these visits happen at a constant rate through the year and time of day, and that the EDs' catchment area is 200,000 people (Williams et al. 2004). This gives a base estimate for the actual demand, λ=10.2 patients per hour. At current average cost of \$200, the demand is given by $\lambda = \Lambda e^{-200 \times 8.5 \times 10^{-4}}$, which gives an estimate of $\Lambda = 12.1$ patients per hour. In our experiments, we vary the catchment area size between 90,000 and 300,000 people, which generates arrival rates that correspond to those of ∼90% trauma hospital EDs in California (OSHPD 2016, ACS 2017).

- The marginal cost of treating a patient at the ED is estimated to be c_o =\$337 in Grannemann et al. (1986) and \$156 in Williams (1996) (both figures inflated to 2016 dollars). We set c_o =\$268 based on these cost figures that is slightly higher than the average \$246.
- To estimate the default ED capacity, μ_o, we once again make use of the observation that the average ED waiting time is two hours. If we assume that the queueing discipline can be approximated by an $M/M/1$ queue then, given the arrival rate of 10.2 patients per hour (see above), we can estimate the average service rate (approximately) $\mu_o = 10.7$ patients per hour. Since we are interested in equilibrium outcomes, in all numerical examples, we report the resulting equilibrium capacity, μ, even if it is less than μ_o.
- We use the following function for the cost of capacity and cost reduction $R(c,\mu) = e^{\beta\mu} + \gamma(c_o - c)^2$, with $\gamma > 0$ and $\beta > 0$. The structure of the first part of the cost function is similar to that used in Grannemann et al. (1986) to estimate the average cost per hospital patient. Since the cost of capacity is largely due to personnel cost, we start from the estimated personnel cost per patient treated at the hospital, which is reported to be to \$110 (Williams (1996), inflated to 2016 prices). At $\mu_o = 10.7$ and $\lambda = 10.2$, we can estimate β by solving $e^{10.7\beta} = 10.2 \times 110$. This produces an estimate of 0.64. In our experiments, we consider the range of β values between 0.44 and 0.84, which corresponds to a cost of capacity per patient of between \$10 and \$784, respectively. Finally, we set $\gamma = 0.054$, which makes the cost of capacity equal to the investment for cost reduction, if all parameters are set to base case. Because the investment in cost reduction is not the focus of this work, we do not perform a detailed sensitivity analysis of the cost-related parameters γ and c_o.

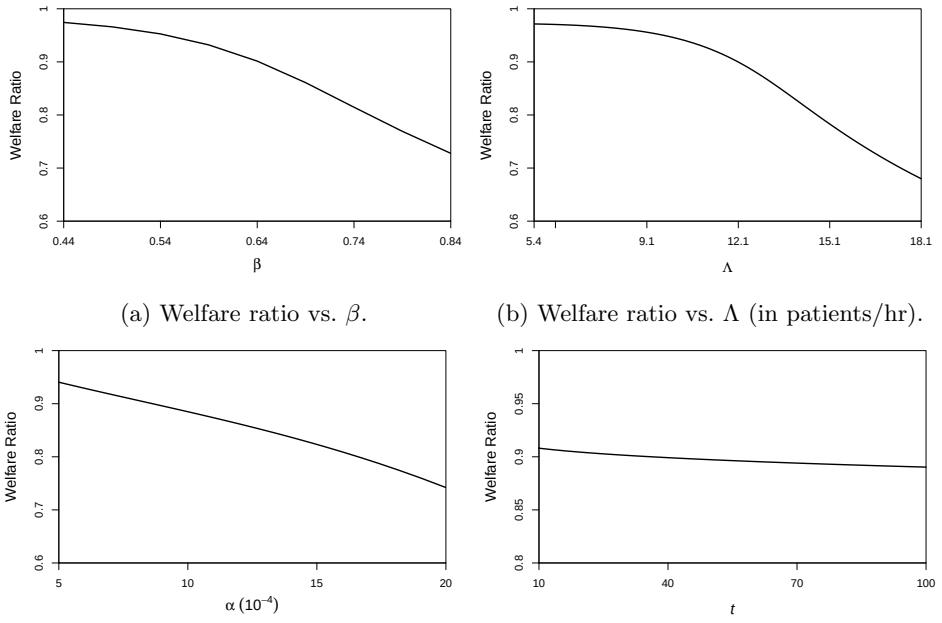
The parameter values chosen, as well as the range within which they are varied (if applicable), are displayed in Table 1. We confirm that for the chosen parameter values, total welfare and the providers' profit functions are concave and the optimal solutions are interior (unless otherwise stated). To maintain connection with reality, we assume that, under second-best regulation, the regulator imposes the \$140 fixed co-payment (see above). This amount is always lower than the optimal first-best price, and our results remain qualitatively similar if the co-payment is reduced to zero.

5.2. Comparison of First-best vs. Second-best Welfare

The loss of welfare associated with implementing second-best yardstick competition, where patients are charged a constant fee, compared to first-best, where patients are charged the welfare-maximizing fee, is presented in Figure 1. A value of 1 indicates that there is no welfare loss. For the parameters tested, we observe that the capacity cost coefficient, β, the size of catchment area, Λ, and the demand elasticity coefficient, α, have the most significant impact on welfare ratio. More specifically, as β, Λ or α increase, the welfare ratio reduces to as low as 68%. The change in cost

Parameter Description	Parameter	Base Estimate	Range
Size of catchment area	Λ	12.1 patient/hr	$[5.4, 18.1]$
Demand elasticity coefficient	α	$9.5 \times 10^{-4}\$^{-1}$	$[5, 20] \times 10^{-4}$
Cost of waiting	t	\$30/hr	$[10, 100]$
Capacity cost coefficient	β	0.64	$[0.44, 0.84]$
Cost-reduction coefficient	γ	0.054	N/A
Default cost per patient	c_o	\$268/patient	N/A

Table 1 Parameter estimates for numerical analysis. The investment cost function is assumed to be $R(c,\mu) = e^{\beta\mu} + \gamma(c_o - c)^2$ and the cumulative distribution function of the patients' benefit from receiving treatment is $\Theta(x) = 1 - e^{-\alpha x}$.



(a) Welfare ratio vs. β.

(b) Welfare ratio vs. Λ (in patients/hr).

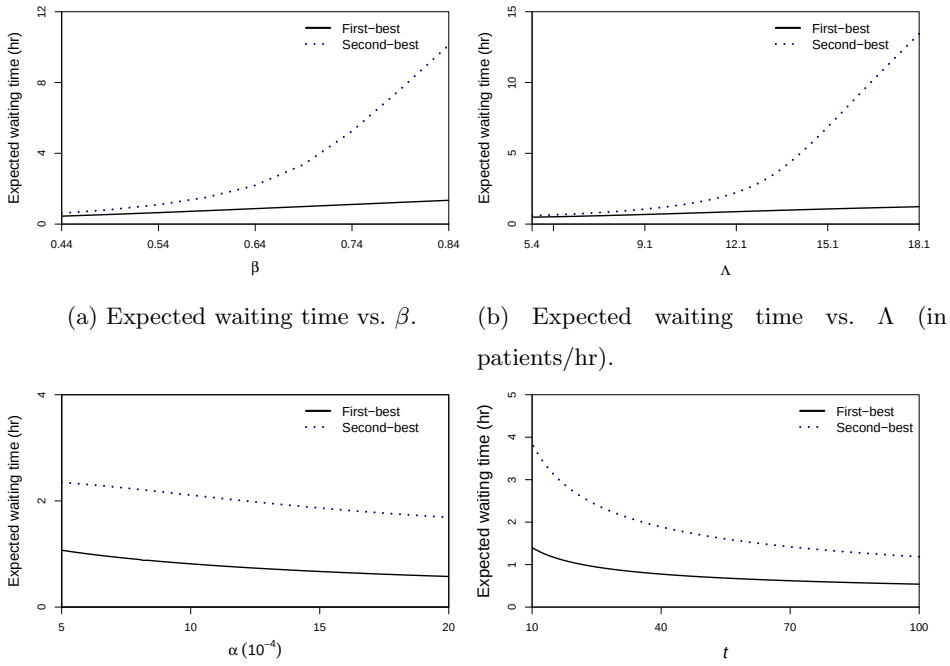
(c) Welfare ratio vs. α.

(d) Welfare ratio vs. t (in \$/hr).

Figure 1 Ratio of second-best to first-best welfare vs. β, Λ, α, and t.

of wait per unit time, t, however, impacts welfare ratio to a much lesser extent. Hence, in situations with relatively large capacity cost, and/or large catchment areas, and/or elastic demand, the additional effort to determine the appropriate fee may be warranted.

We next investigate what drives the impact of each of the four parameters, β, Λ, α, and t, on the welfare ratio described above. We start with the capacity cost coefficient β. As β increases, capacity becomes more costly, therefore providers choose to operate at higher utilization levels (defined as the effective arrival rate divided by the capacity) under both first- and second-best regulation, resulting in average waiting times that are increasing in β – see Figure 2(a). We note that waiting times increase more for second-best as opposed to first-best regulation. This is due to the fact that customers over-join under second-best regulation, coupled with the fact that expected waiting times become more sensitive to increases in arrival rate in an $M/M/1$ queue as it becomes



(a) Expected waiting time vs. β.

(b) Expected waiting time vs. Λ (in patients/hr).

(c) Expected waiting time vs. α.

(d) Expected waiting time vs. t (in \$/hr).

Figure 2 Expected waiting time (in hours) under first-best and second-best regulation vs. β, Λ, α, and t.

more congested. Hence, as β gets larger, the welfare loss associated with second-best regulation increases. We observe a similar phenomenon as Λ increases, see Figure 2(b).

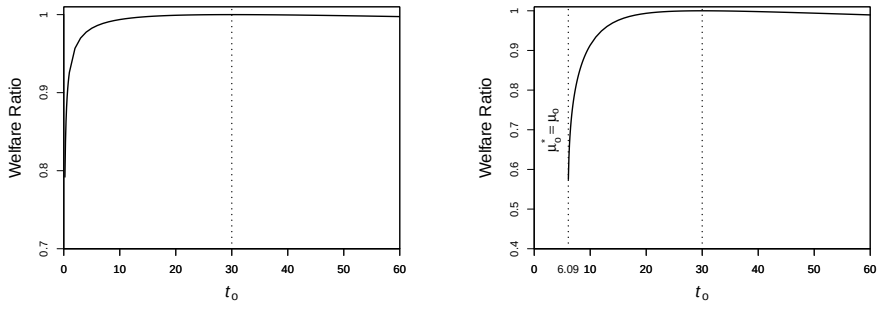
As the demand sensitivity parameter α decreases, demand becomes less price/wait-time sensitive and so more patients are willing to visit the ED for any given price/wait time. Under first-best regulation, as α decreases, the regulator can react to the associated increase in demand using two levers: i) increase capacity in order to serve the increased demand faster; ii) increase the price to curtail the increase in arrivals. In our numerical analysis, we find that the regulator will increase both capacity and price as α decreases. Nevertheless, the increase in capacity under first-best regulation is not enough to reduce waiting times which will increase as α decreases, see Figure 2(c). In contrast, under second-best regulation, as demand becomes less price-sensitive (i.e., α decreases), the regulator only has the first lever available; he can increase capacity but cannot charge a higher price. Furthermore, increasing capacity is more effective in increasing social welfare as arrivals increase, that is, when α is lower. As a result, we observe that, as α decreases under second-best regulation, waiting times also increase, but the gap between the waiting times under first- and second-best regulation remains roughly constant, see Figure 2(c). Naturally, since the gap in waiting times remains constant as α increases while the total welfare decreases as demand becomes more price-sensitive (i.e., α increases), the welfare loss associated with second-best compared to first-best regulation also increases in α, as observed in Figure 1.

We next examine the impact of the cost of waiting t on the welfare ratio. We note that it has

a less pronounced impact on the welfare ratio than the other parameters, β, Λ, or α, as shown in Figure 1. To see why this is the case, note that if t is high, then under either first- or second-best regulation, the system operates at relatively low utilization, resulting in relatively low waiting times, see Figure 2(d). Hence, the over-joining behavior observed under second-best regulation does not affect social welfare as much. If, on the other hand, t is low, under either type of regulation, the system will operate under high utilization, resulting in long waiting times, see Figure 2(d). However, since the waiting time cost t is low, customers are less sensitive to delays and therefore social welfare is, again, not greatly affected by the inefficient over-joining behavior under second-best regulation.

5.3. Impact of Misestimating the Cost of Waiting

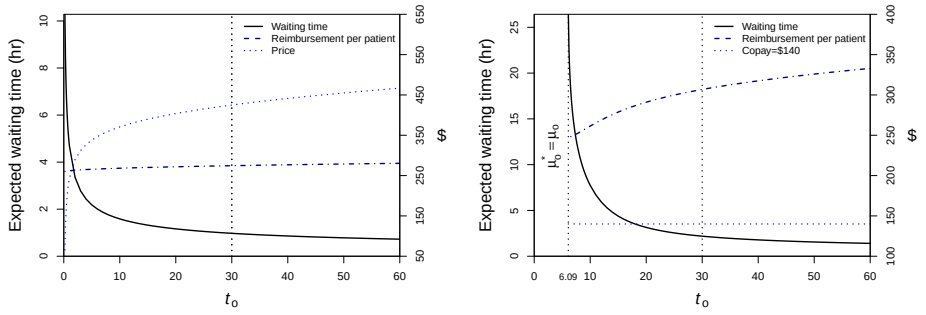
To implement the proposed payment schemes, the regulator needs to estimate the cost of waiting, t. We next investigate the impact of misestimation of t on welfare using the following procedure. We assume that $t = \$30$ and that the regulator erroneously sets this cost equal to $t_o (\neq t)$ in (16)-(17) for first-best and in (23) for second-best regulation. We identify the equilibrium for each t_o ranging from \$0.1 to \$60 in increments of \$0.1 by solving the FOCs of the provider's objective and verifying that the objective is maximized with these actions. We then compare the welfare in this equilibrium to the base case, that is, when the regulator estimates t correctly. We present the welfare ratio as a function of t_o in Figures 3(a) and (b) for first- and second-best regulation, respectively. Similarly, Figures 4(a) and (b) present the resulting equlibrium waiting times, regulator reimbursement, and patient price (or co-payment). Under second-best regulation, the providers exert no effort in capacity expansion in equilibrium for $t_o \leq \$6.09$, hence we plot the welfare ratio beyond this point only.



(a) First-best regulation. (b) Second-best regulation.

Figure 3 Ratio of realized to first-best (second-best) welfare vs. misestimated cost of waiting t_o. Actual cost of waiting is $= \$30$. Other parameters set to base case.

It is clear from Figure 3 that estimation error in the cost of waiting, t, generates a loss of welfare. When t is underestimated, providers operate under high utilization simply because they lack sufficient incentives to cut the high wait times that can be seen in Figure 4. When t is



(a) First-best regulation. (b) Second-best regulation.

Figure 4 Expected waiting time, price or co-payment, and reimbursement per patient vs. misestimated cost of waiting t_o. Actual cost of waiting is $t = \$30$. Other parameters set to base case.

overestimated, the providers invest more in costly capacity, which is underutilized leading to short waiting times. Nevertheless, there are two interesting observations. First, the welfare loss under first-best regulation is less sensitive to estimation errors than under second-best. Second, the impact of the estimation error is more substantial when t is underestimated (it reaches 80% and 54% in first- and second-best, respectively) compared to when it is overestimated (it is above 99.5% and 97.7% in first- and second-best, respectively, when $t_o \geq t$).[6] We believe that this has to do with the fact that waiting times are convex in capacity – starting from the optimal capacity (set when t is estimated accurately), a small decrease in capacity (due to t being underestimated), will generate a greater loss of welfare due to a larger increase in waiting times than the welfare gain associated with the small reduction in waiting times generated by a small increase in capacity (due to t being overestimated). Hence, the welfare loss is more substantial when t is underestimated than when it is overestimated.

The fact that total welfare is not very sensitive to the actual cost of waiting assumed by the regulator points to the fact that welfare is relatively flat around the actual cost, t. An overestimate of t will generate a welfare loss due to installing more capacity than optimal. This will be almost fully compensated by an increase in welfare due to the associated reduction in waiting times and increase in demand. In light of this discussion, the regulator may be able to use the waiting-time cost parameter, t, as a lever to influence waiting times; by choosing to implement a regulatory scheme with high t, the regulator shifts the equilibrium outcome towards lower waiting times, at the expense of higher hospital costs, without sacrificing much in terms of welfare. We quantify

[6] We verify the robustness of this observation by generating 1,000 random scenarios for Λ, α, and β using the ranges specified in Table 1. In all scenarios, we set $t = \$30$. As in the case above, we find the welfare loss under first- and second-best regulation if the regulator first underestimates t to $t_o = \$20$ and second, overestimates t to $t_0 = \$40$. In all the parameter combinations, the welfare loss for first-best regulation was minimal; the average welfare loss was equal to 0.10%, with maximum loss equal to 0.21% for $t_o = \$20$, and it was equal to 0.05% with maximum loss equal to 0.11%, for $t_0 = \$40$. In the second-best for $t_o = \$20$, the average welfare loss was equal to 1.80% with maximum loss equal to 6.78%, and for $t_o = \$40$, the average welfare loss was 0.75% with maximum loss equal to 2.68%.

these observations in Figure 4. If the regulator uses the second-best yardstick competition scheme and sets penalties assuming (correctly) that $t_o = t = 30$, the average waiting time is 2.18 hours, while the reimbursement per patient is \$306. By increasing the penalty to $t_o = 40$ the waiting time can be reduced to 1.78 hours on average at a cost of \$317 per patient. This change entails less than 2% loss in welfare.

6. Extensions

In the previous sections we assumed that providers were acting as local monopolists. In this section we summarize the results of an extension where providers compete for customers based on waiting time and, where applicable, prices. Due to space restrictions, the full extension is presented in §EC.1 of this paper. We note below the main observations from this analysis. First, competition does not provide enough incentives for providers to invest optimally in capacity or cost reduction. Therefore, some regulatory intervention is warranted. Second, the standard cost-based yardstick competition is still ineffective in incentivizing capacity investment, even when providers compete based on waiting times. This happens because, in equilibrium, the marginal value of additional customers for each provider is zero, therefore providers have no incentive to increase capacity. Third, the first-best yardstick competition proposed in §4.3 for the monopoly setting still achieves first-best outcomes in the presence of demand-side competition. Fourth, the simple second-best scheme proposed in §4.4 – which reimburses providers based on their relative waiting-time performance – fails to incentivize capacity investment in the presence of direct competition. This is because competition renders the waiting-time benchmarks irrelevant – due to competition all (active) providers have the same waiting time, irrespective of how much capacity they have installed. Nevertheless, we show that there exits a relatively straightforward modification, based on dividing providers into disjoint competing and non-competing sets, that restores second-best outcomes.

Furthermore, we are able to show that the model can be extended in a number of directions (see §EC.2). Namely, we examine the case of multiple customer classes, time-varying arrivals, more general cost structure, regulation based on tail-statistics instead of average waiting time, provider heterogeneity, exogenous arrivals, and more general queueing models such as Jackson networks.

7. Conclusions

This paper investigates the use of yardstick competition, a regulatory scheme that creates cost-reduction incentives (Shleifer 1985), in service settings where, in addition to cost control, the regulator is also interested in incentivizing wait-time reduction. This scheme has proliferated in the regulation (and reimbursement) of hospitals (Fetter 1991). As we summarize in Table 2, we find that the standard form of yardstick competition fails in this second dimension of performance.

Perhaps this finding helps explain the persistently long waiting times experienced by patients in many healthcare systems.

We also present two alternative schemes that fare better. The first scheme, which involves a provider-specific customer fee, achieves first-best investment in both cost and wait-time reduction, but is rather difficult to implement in practice – besides the customer fee being politically sensitive in the healthcare setting, this scheme places a high informational burden on the regulator with respect to the queueing discipline. The second scheme, which assumes that the service is funded exclusively through transfer payments (e.g., taxes or insurance premia), may be easier to implement. In essence, this scheme modifies the transfer payment of the standard cost-based yardstick competition by adding a component which is decreasing in the difference between the average waiting times of each provider and that of an exogenous benchmark constructed by averaging the waiting time of all other providers. The simplicity of this second regulatory scheme comes at a cost of efficiency as it no longer achieves first-best incentives. Nevertheless, as our numerical investigation illustrates, it is likely to be better than the status quo where waiting-time reduction is not incentivized.

We hope that this paper will contribute to the current debate on how to best incentivize investment in waiting-time reduction in healthcare, particularly in EDs where waiting times have been argued to be undesirably long. In fact, our paper provides a high-level guideline for regulators, such as CMS in the US and the National Health Service (NHS) in the UK who have started monitoring ED waiting times, on how to use waiting-time information in the reimbursement formula. We believe that this is a promising alternative to top-down targets, such as the four-hour target that has been implemented in the UK for patients visiting EDs (see, e.g., Siddique (2016)).

Of course, the exact application may be complicated, especially by concerns about patient selection based on service times or system congestion. We believe this may not be a problem in practice, as was the case with the advent of cost-based yardstick competition which is not believed to have given rise to significant selection based on costs. Nevertheless, understanding and mitigating selection problems that arise in the presence of waiting-time yardstick competition is an issue that future research should address. An additional limitation of this work is that all of the schemes proposed assume that the regulator knows the cost of customer waiting per unit time, t. This may not always be the case, but, as we show in our numerical investigation, total welfare is not sensitive to the precise value that it takes. In fact, one may view the waiting-time cost parameter, t, as a lever that can be used to influence waiting times; by choosing to implement a regulatory scheme with high t, the regulator shifts the equilibrium outcome towards lower waiting times at the expense of higher costs, with little loss in overall welfare. Nevertheless, identifying a modified scheme that does not require this information may be a promising direction for further research.

Reimbursement Scheme (Section Presented)	Customer Payment	Transfer Payment	Incentives to Reduce Cost	Incentives to Reduce Waiting Time	Informational Requirement
Cost-of-Service (§4.1)	Cost of service.	Set equal to investment cost.	No	No	Ex post costs.
Cost-based Yardstick Competition (§4.2)	Average cost of service at other providers.	Set equal to investment cost of other providers.	Yes	No	Ex post costs of all providers.
Cost- and Capacity-based Yardstick Competition (§4.3)	Average cost of service plus an additional toll equal to the waiting-time externality.	Set equal to investment cost of other providers minus the toll paid by customers, plus two additional terms: i) a term proportional to the difference between the industry average costs (excluding the focal provider) and the costs of the focal provider; and ii) a term proportional to the difference in the industry average service rate (excluding the focal provider) and that of the focal provider.	Yes (First Best)	Yes (First Best)	Ex post costs, arrival rates, and service rates of all providers; derivatives of waiting time function with respect to service rate and arrival rate.
Free-at-the-point-of-care Yardstick Competition (§4.4)	Zero (or an exogenous constant).	Set equal to investment and service cost of other providers, with an additional term proportional to the difference in waiting time at the focal provider compared to industry average (excluding the focal provider).	Yes (Second Best)	Yes (Second Best)	Ex post costs, arrival rates, and waiting times of all providers.

Table 2: The summary of reimbursement schemes analyzed in §§4.1–4.4.

Finally, we note that our analysis assumed that service providers are regional monopolists or compete perfectly (see §EC.1). It would be of interest to examine the performance of the proposed schemes under imperfect (horizontally differentiated) competition.

Acknowledgments

The authors are grateful to Sergeui Netessine (department editor), the anonymous associate editor, and three anonymous referees for comments that have greatly improved the paper. The authors acknowledge funding from the Institute of Innovation and Entrepreneurship of the London Business School (`https://www.london.edu/faculty-and-research/research-centres/iie`).

References

ACS. 2017. Trauma Centers. Accessed September 27, 2017, `https://www.facs.org/search/trauma-centers?state=CA`.

Adida, E., H. Mamani, S. Nassiri. 2016. Bundled payment vs. fee-for-service: Impact of payment scheme on performance. *Management Science* **63**(5) 1606–1624.

Afeche, P. 2013. Incentive-compatible revenue management in queueing systems: Optimal strategic delay. *Manufacturing & Service Operations Management* **15**(3) 423–443.

Akan, M., B. Ata, M. A. Lariviere. 2011. Asymmetric information and economies of scale in service contracting. *Manufacturing & Service Operations Management* **13**(1) 58–72.

Allon, G., A. Federgruen. 2008. Service competition with general queueing facilities. *Operations Research* **56**(4) 827–849.

Andritsos, D. A., S. Aflaki. 2015. Competition and the operational performance of hospitals: The role of hospital objectives. *Production and Operations Management* **24**(11) 1812–1832.

Andritsos, D. A., C. S. Tang. 2015. Incentive programs for reducing readmissions when patient care is co-produced. Working Paper No. MOSI-2015-1110, HEC Paris, Paris, France.

Armstrong, M., R. H. Porter. 2007. *Handbook of Industrial Organization Vol. 3*. Elsevier, Amsterdam, Netherlands.

Bakshi, N., S. H. Kim, N. Savva. 2015. Signaling new product reliability with after-sales service contracts. *Management Science* **61**(8) 1812–1829.

Batt, R. J., C. Terwiesch. 2015. Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Science* **61**(1) 39–59.

Brown, A. M., S. L. Decker, F. W. Selck. 2015. Emergency department visits and proximity to patients' residences, 2009-2010. *NCHS Data Brief* (192) 1–8.

Bureau of Labor Statistics. 2016. May 2016 State Occupational Employment and Wage Estimates. Accessed September 18, 2017, `https://www.bls.gov/oes/current/oessrcst.htm#top`.

Cachon, G., P. T. Harker. 2002. Competition and outsourcing with scale economies. *Management Science* **48**(10) 1314–1333.

Cachon, G. P., F. Zhang. 2006. Procuring fast delivery: Sole sourcing with information asymmetry. *Management Science* **52**(6) 881–896.

CEB. 2016. Medical plan trends and observations report. Accessed December 11, 2016, https://www.cebglobal.com/human-resources/total-rewards/medical-plan-trends.html.

CIHI. 2012. Canadians continue to wait for care. Accessed December 11, 2016, https://www.cihi.ca/en/health-system-performance/access-and-wait-times/canadians-continue-to-wait-for-care.

CMS. 2011. National hospital ambulatory medical care survey: 2011 emergency department summary tables. Accessed December 11, 2016, http://www.cdc.gov/nchs/data/ahcd/nhamcs_emergency/2011_ed_web_tables.pdf.

CMS. 2014. Financial report fiscal year 2014. Accessed December 11, 2016, https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/CFOReport/Downloads/CMS-Financial-Report-for-Fiscal-Year-2014.pdf.

CMS. 2016a. Hospital value-based purchasing. Accessed December 11, 2016, https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/downloads/Hospital_VBPurchasing_Fact_Sheet_ICN907664.pdf.

CMS. 2016b. Hospital Compare: Find a hospital. Accessed December 11, 2016, https://www.medicare.gov/hospitalcompare/search.html.

Dada, M., W.D. White. 1999. Evaluating financial risk in the medicare prospective payment system. *Management Science* **45**(3) 316–329. doi:10.1287/mnsc.45.3.316.

Dai, T., M. Akan, S. Tayur. 2016. Imaging room and beyond: The underlying economics behind physicians test-ordering behavior in outpatient services. *Manufacturing & Service Operations Management* **19**(1) 99–113.

Deo, S., I. Gurvich. 2011. Centralized vs. decentralized ambulance diversion: A network perspective. *Management Science* **57**(7) 1300–1319.

Edelson, N. M., D. K. Hilderbrand. 1975. Congestion tolls for Poisson queuing processes. *Econometrica* **43**(1) 81–92.

Ellis, R. P., T. G. McGuire. 1986. Provider behavior under prospective reimbursement: Cost sharing and supply. *Journal of Health Economics* **5**(2) 129–151.

Fetter, R. B. 1991. Diagnosis related groups: Understanding hospital performance. *Interfaces* **21**(1) 6–26.

Freeman, M., N. Savva, S. Scholtes. 2016. Economies of scale and scope in hospitals. Working paper, Judge Business School, University of Cambridge, Cambridge, UK.

GAO. 2009. Hospital emergency departments: Crowding continues to occur, and some patients wait longer than recommended time frames. Accessed December 11, 2016, http://www.gao.gov/new.items/d09347.pdf.

Gaynor, M. 2004. Competition and quality in hospital markets. What do we know? What don't we know? *Economie Publique,* **15**(2) 3–40.

Grannemann, T. W., R. S. Brown, M. V. Pauly. 1986. Estimating hospital costs: A multiple-output analysis. *Journal of Health Economics* **5**(2) 107–127.

Guo, P., C. S. Tang, Y. Wang, M. Zhao. 2016. The impact of reimbursement policy on patient welfare, readmission rate and waiting time in a public healthcare system: Fee-for-service vs. bundled payment. Working paper, Anderson School of Management, University of California, Los Angeles, CA.

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

Hassin, R. 2016. *Rational Queueing*. CRC Press, Boca Raton, FL.

Hassin, R., M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers, Norwell, MA.

Holmstrom, B. 1982. Moral hazard in teams. *The Bell Journal of Economics* **13**(2) 324–340.

Jamasb, T., M. Pollitt. 2000. Benchmarking and regulation: International electricity experience. *Utilities Policy* **9**(3) 107–130.

Jiang, H., Z. Pang, S. Savin. 2016. Capacity management for outpatient medical services under competition and performance-based incentives. Working paper, Wharton School, University of Pennsylvania, Philadelphia, PA.

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

Kim, S. H., M. A. Cohen, S. Netessine, S. Veeraraghavan. 2010. Contracting for infrequent restoration and recovery of mission-critical systems. *Management Science* **56**(9) 1551–1567.

Kleinrock, L. 1975. *Queueing Systems, Volume I: Theory*. John Wiley & Sons Inc., New York, NY.

Laffont, J. J., J. Tirole. 1993. *A Theory of Incentives in Procurement and Regulation*. MIT Press, Cambridge, MA.

Lariviere, M. A., J. A. Van Mieghem. 2004. Strategically seeking service: How competition can generate poisson arrivals. *Manufacturing & Service Operations Management* **6**(1) 23–40.

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

Lee, H. L., M. A. Cohen. 1985. Multi-agent customer allocation in a stochastic service system. *Management Science* **31**(6) 752–763.

Lefouili, Y. 2015. Does competition spur innovation? The case of yardstick competition. *Economics Letters* **137** 135–139.

Ma, C. A. 1994. Health care payment systems: Cost and quality incentives. *Journal of Economics & Management Strategy* **3**(1) 93–112.

Mayes, R. 2007. The origins, development, and passage of Medicare's revolutionary prospective payment system. *Journal of the History of Medicine and Allied Sciences* **62**(1) 21–55.

Mendelson, H., S. Whang. 1990. Optimal incentive-compatible priority pricing for the M/M/1 queue. *Operations research* **38**(5) 870–883.

Meran, G., C. Von Hirschhausen. 2009. A modified yardstick competition mechanism. *Journal of Regulatory Economics* **35**(3) 223–245.

Nalebuff, B. J., J. E. Stiglitz. 1983. Information, competition, and markets. *The American Economic Review* **73**(2) 278–283.

Naor, P. 1969. The regulation of queue size by levying tolls. *Econometrica* **37**(1) 15–24.

OSHPD. 2016. Emergency Department and Ambulatory Surgery Data. Accessed September 18, 2017, `https://oshpd.ca.gov/HID/ED-AS-Data.html#Encounters`.

Pope, G. C. 1989. Hospital nonprice competition and Medicare reimbursement policy. *Journal of Health Economics* **8**(2) 147–172.

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

Ringel, J. S., S. D. Hosek, B. A. Vollaard, S. Mahnovski. 2002. The elasticity of demand for health care. A review of the literature and its application to the military health system. Tech. rep., National Defense Research Institute, RAND Health, Santa Monica, CA.

Roques, F. A., N. Savva. 2009. Investment under uncertainty with price ceilings in oligopolies. *Journal of Economic Dynamics and Control* **33**(2) 507–524.

Saghafian, S., G. Austin, S. J. Traub. 2015. Operations research/management contributions to emergency department patient flow optimization: Review and research prospects. *IIE Transactions on Healthcare Systems Engineering* **5**(2) 101–123.

Saghafian, S., W. J. Hopp, M. P. Van Oyen, J. S. Desmond, S. L Kronick. 2014. Complexity-augmented triage: A tool for improving patient safety and operational efficiency. *Manufacturing & Service Operations Management* **16**(3) 329–345.

Sawkins, J. W. 1995. Yardstick competition in the English and Welsh water industry Fiction or reality? *Utilities Policy* **5**(1) 27–36.

Shleifer, A. 1985. A theory of yardstick competition. *The RAND Journal of Economics* **16**(3) 319–327.

Siddique, H. 2016. Hospital A&E waiting times in England rise by a third in November. *The Guardian* (January 14), `https://www.theguardian.com/society/2016/jan/14/hospital-waiting-times-england-rise-november-accident-emergency`.

So, K. C., C. S. Tang. 2000. Modeling the impact of an outcome-oriented reimbursement policy on clinic, patients, and pharmaceutical firms. *Management Science* **46**(7) 875–892.

Sobel, J. 1999. A reexamination of yardstick competition. *Journal of Economics & Management Strategy* **8**(1) 33–60.

Tangerås, T. P. 2009. Yardstick competition and quality. *Journal of Economics & Management Strategy* **18**(2) 589–613.

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

Williams, B., J. Nicholl, J. Brazier. 2004. Accident and emergency departments. A. Stevens, J. Raftery, eds., *Health Care Needs Assessment: The Epidemiologically Based Needs Assessment Reviews*. Radcliffe Medical Press, Oxford, UK. 1-54.

Williams, R. M. 1996. Distribution of emergency department costs. *Annals of Emergency Medicine* **28**(6) 671–676.

Zhang, D. J., I. Gurvich, J. A. Van Mieghem, E. Park, R. S. Young, M. V. Williams. 2016. Hospital readmissions reduction program: An economic and operational analysis. *Management Science* **62**(11) 3351–3371.

Zorc, S., S.E. Chick, S. Hasija. 2017. Outcomes-based reimbursement policies for chronic care pathways Working Paper No. 2017/35/DSC/TOM, INSEAD, Fontainebleau, France.

Appendix

Proof of Proposition 1: Under the assumption that FOCs are necessary and sufficient to obtain the first-best outcomes (see Appendix EC.3 for conditions that guarantee that this is the case), the first-best price, p^*, marginal cost, c^*, and capacity, μ^*, are the unique solutions to $\frac{\partial}{\partial p}S(p,c,\mu)=0$, $\frac{\partial}{\partial c}S(p,c,\mu)=0$ and $\frac{\partial}{\partial \mu}S(p,c,\mu)=0$, which yield (10)–(12). The first-best transfer payment, T^*, is obtained by solving for $\Pi(c^*,\mu^*|p^*,T^*)=0$, which leads to (13) by (4). □

Proof of Proposition 2: See proof of Proposition 1 in Shleifer (1985).

Proof of Proposition 3: Assume that the following sufficient condition in Shleifer (1985) holds (see §EC.3.2 for details)

$$\frac{\partial \lambda(c,\mu_o)}{\partial c}+\frac{\partial^2 R(c,\mu_o)}{\partial c^2}>0. \tag{25}$$

For the regulatory scheme given in Proposition 1, by (4) provider i's profit function is

$$\Pi(c_i,\mu_i|p_i,T_i)=(\bar{c_i}-c_i)\lambda(\bar{c_i},\mu_i)-R(c_i,\mu_i)+\bar{R}_i. \tag{26}$$

Because $\frac{\partial}{\partial \mu_i}\Pi(c_i,\mu_i|p_i,T_i)=(\bar{c_i}-c_i)\frac{\partial}{\partial \mu_i}\lambda(\bar{c_i},\mu_i)-\frac{\partial}{\partial \mu_i}R(c_i,\mu_i)$, in any symmetric equilibrium where $\bar{c_i}=c_i$, we have $\frac{\partial}{\partial \mu_i}\Pi(c_i,\mu_i|p_i,T_i)<0$ for all $\mu_i\geq\mu_o$. Thus, in all potential symmetric equilibria, all providers choose their default capacity level, μ_o. Also, because $R(c,\mu_o)$ is convex, $\Pi(c_i,\mu_o|p_i,T_i)$ is concave in c_i. By (4), provider i's optimal marginal cost is obtained by

$$\frac{\partial}{\partial c_i}\Pi(c_i,\mu_o|p_i,T_i)=-\lambda(\bar{c_i},\mu_o)-\frac{\partial}{\partial c_i}R(c_i,\mu_o)=0, \tag{27}$$

which holds at a unique $c_i=\bar{c_i}=\check{c}$. Hence, there exists a unique symmetric equilibrium where all providers choose capacity level μ_o and marginal cost level $\check{c}$ (and make zero profit). We next show that $\check{c}>c^*$ under the additional assumptions that $\frac{\partial^2 R(c,\mu)}{\partial c\partial \mu}\geq 0$ and (25) holds for all $\mu\geq\mu_o$. If $\frac{\partial^2 R(c,\mu)}{\partial c\partial \mu}\geq 0$, because $\lambda(p,\mu)$ is strictly increasing in μ by (2), $\frac{\partial}{\partial \mu}W(\lambda,\mu)<0$ and $\frac{\partial}{\partial \lambda}W(\lambda,\mu)>0$, $\left(\lambda(c,\mu)+\frac{\partial R(c,\mu)}{\partial c}\right)$ is strictly increasing in μ for $c\in(0,c_o]$. Thus, for $\mu^*>\mu_o$, we have

$$\lambda(\check{c},\mu^*)+\frac{\partial R(\check{c},\mu^*)}{\partial c}>\lambda(\check{c},\mu_o)+\frac{\partial R(\check{c},\mu_o)}{\partial c}=0, \tag{28}$$

where the equality follows from the fact that $\check{c}$ satisfies (27) by definition in the unique symmetric equilibrium. By (25) if $c^*\geq\check{c}$, then

$$\lambda(c^*,\mu^*)+\frac{\partial R(c^*,\mu^*)}{\partial c}\geq\lambda(\check{c},\mu^*)+\frac{\partial R(\check{c},\mu^*)}{\partial c},$$

which, along with (28), leads to $\lambda(c^*,\mu^*)+\frac{\partial R(c^*,\mu^*)}{\partial c}>0$. However, this contradicts the optimality of (c^*,μ^*) in the welfare maximization problem because (10) cannot hold. Thus $c^*<\check{c}$. □

Proof of Theorem 1: Assume that the FOCs are necessary and sufficient to obtain the optimal actions of each provider (see §EC.3.3 for sufficient conditions). If the regulator sets service provider i's price equal to p_i given in (16) and transfer payment equal to T_i given in (17), provider i's objective function is as given in (18) by (4). We next show that there is a unique symmetric equilibrium. Let $a_j=(\tilde{c},\tilde{\mu})$ denote the action of provider j for all $j\neq i$ and let $\tilde{\lambda}$ denote the associated arrival rate that satisfies (2) with price set as in (16).

By (18), the FOCs of Π for provider i are

$$\frac{\partial}{\partial c}\Pi(c_i,\mu_i)=-\tilde{\lambda}-\frac{\partial}{\partial c}R(c_i,\mu_i)=0, \tag{29}$$

$$\frac{\partial}{\partial \mu}\Pi(c_i,\mu_i)=-t\frac{\partial}{\partial \mu}W(\tilde{\mu},\tilde{\lambda})\tilde{\lambda}-\frac{\partial}{\partial \mu}R(c_i,\mu_i)=0. \tag{30}$$

If $\tilde{c} = c^*$ and $\tilde{\mu} = \mu^*$, then $\tilde{\lambda} = \lambda^*$ by (2) and (16). Because (10)–(12) have a unique solution, so do (29)–(30). In addition, because FOCs are necessary and sufficient to obtain the optimal actions of each provider, (c^*, μ^*) is a Nash equilibrium. It is easy to check that providers make zero profit in equilibrium.

Now consider any other $\tilde{c}$ and $\tilde{\mu}$. In order for provider i to pick the same actions, $\tilde{c}$ and $\tilde{\mu}$ have to satisfy the FOCs (29) and (30) because the FOCs are necessary and sufficient to obtain the optimal actions of each provider. However, because S has a unique optimal solution and the FOCs are sufficient, for $\tilde{c}$ and $\tilde{\mu}$ to be a solution to (29) and (30), they must satisfy $\tilde{c} = c^*$ and $\tilde{\mu} = \mu^*$. Hence, (c^*, μ^*) is the unique symmetric equilibrium. □

Proof of Theorem 2: Assume that the FOCs are necessary and sufficient to obtain the optimal actions of each provider (see § EC.3.5 for sufficient conditions). Assume that the regulator pays the transfer payment T_i defined as in (23) to provider i, for $i = 1, \ldots, N$, and customers are not charged a toll. The proof of the result is similar to that of Theorem 1.

When patients are not charged a toll, the objective of the regulator is

$$S(c, \mu) = \Lambda \int_{tW(\lambda(\mu),\mu)}^{\infty} (x - tW(\lambda(\mu), \mu))\, d\Theta(x) - c\lambda(\mu) - R(c, \mu). \tag{31}$$

By Leibniz rule $\frac{\partial}{\partial y}\left(\Lambda \int_{\bar{\Theta}^{-1}\left(\frac{y}{\Lambda}\right)}^{\infty} x d\Theta(x)\right) = \bar{\Theta}^{-1}\left(\frac{y}{\Lambda}\right)$ for $y \in [0, \Lambda]$. Hence, by (2), the FOCs of $S(c, \mu)$ are given by (19) and (20). Because FOCs are assumed to be necessary and sufficient, (19) and (20) have a unique solution, which is μ_o^* and c_o^*.

We next show that $\mu_i = \mu_o^*$ and $c_i = c_o^*$ for $i = 1, \ldots, N$ is an equilibrium under the scheme given in Theorem 2. Assume that each provider, except provider i, picks $(\tilde{c}, \tilde{\mu})$. Then, by (24) provider i's optimal actions satisfy

$$\frac{\partial}{\partial c}\Pi(c, \mu) = -\lambda(\mu) - \frac{\partial}{\partial c}R(c, \mu) = 0, \tag{32}$$

$$\frac{\partial}{\partial \mu}\Pi(c, \mu) = -c\lambda'(\mu) - t\tilde{\lambda}\frac{\partial}{\partial \mu}W(\lambda(\mu), \mu) - \frac{\partial}{\partial \mu}R(c, \mu) = 0, \tag{33}$$

because the FOCs are necessary and sufficient to obtain the optimal actions of each provider. If $\tilde{c} = c_o^*$ and $\tilde{\mu} = \mu_o^*$, because (19) and (20) have a unique solution (c_o^*, μ_o^*), so do (32) and (33). Because (c_o^*, μ_o^*) is the solution to (19) and (20), (c_o^*, μ_o^*) is a symmetric equilibrium. It can easily be shown that providers make zero profit in this equilibrium. Uniqueness of the symmetric equilibrium follows as in the proof of Theorem 1. □