Electronic Companion:
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

The electronic companion (EC) presents an extension of the main model to the case where providers compete directly for customers in §EC.1. §EC.2 focuses on the free-at-the-point-of-care (second-best) yardstick-competition scheme of §4.4 of the main paper and presents extensions to multiple customer classes, time-varying arrival rates, more general cost structure, using tail-statistics instead of average waiting time, provider heterogeneity, purely exogenous arrivals, and more general queueing models. §EC.3 examines second-order conditions for the providers' and the regulator's objective functions that ensure that first-order conditions (FOCs) are sufficient to guarantee optimality. §EC.4 and §EC.5 present alternative payment schemes that achieve first- and second-best equilibria, respectively, for which we can prove that the equilibrium is unique (i.e., rule out the existence of asymmetric equilibria).

EC.1. Yardstick Regulation When Providers Compete

In some settings, service providers have to compete with each other for customers. For example, urbanized areas may have multiple Emergency Departments (EDs) and patients may choose which ED to visit based on information regarding waiting times. Such choice has been enhanced by online tools that publish average ED waiting times (CMS 2016, ProPublica 2015). In this section, we extend the model to allow providers to compete directly for customers based on waiting times and, where applicable, on prices. This turns the analysis into a more complex bi-level game, where customers choose service providers, and service providers choose their level of capacity and cost-reduction effort.

More specifically, we assume that $N \geq 2$ identical competing providers exist that serve a single catchment area. The notation and model in this section are similar to those introduced in Section 3 with the following straightforward extensions: the prices, capacity, and cost per patient at each provider are given by $p=(p_1,p_2,\ldots,p_N)$, $\mu=(\mu_1,\mu_2,\ldots,\mu_N)$, $c=(c_1,c_2,\ldots,c_N)$, respectively, and the vector $\Sigma=(c,\mu,p)$ summarizes all this information. Therefore, the customer arrival rate at provider i in equilibrium will be given by $\lambda_i(\Sigma)$, and let $\lambda=(\lambda_1,\lambda_2,\ldots,\lambda_N)$ and $\Xi(\Sigma)=\sum_{i=1}^N \lambda_i(\Sigma)$.

We start by finding the equilibrium joining decisions of the customers. As in §3.1, service needs arise through a stochastic process with rate Λ, and each customer with a service need may decide to seek service or not. However, now he may choose among N providers and we assume that he chooses the provider with the lowest total cost, as long as this cost is less than the benefit of receiving the service, r. For simplicity, we ignore any differences in travel costs incurred by the customer when choosing which provider to visit. In equilibrium, the individual arrival rates to each provider $\lambda_i(\Sigma)$ must satisfy the following conditions (see also Chen and Wan (2005), Afanasyev and Mendelson (2010) for similar models):

$$\text{If } p_i+tW(0,\mu_i)>p_j+tW(\lambda_j(\Sigma),\mu_j) \text{ for some } j, \text{ then } \lambda_i(\Sigma)=0, \tag{EC1}$$

$$\text{If } \lambda_i(\Sigma)>0 \text{ and } \lambda_j(\Sigma)>0 \text{ for } i,j=1,\ldots,N \text{ then } p_i+W(\lambda_i(\Sigma),\mu_i)=p_j+W(\lambda_j(\Sigma),\mu_j), \tag{EC2}$$

$$\Xi(\Sigma)=\Lambda\bar{\Theta}\left(\min_{i=1,\ldots,N}\{p_i+tW(\lambda_i(\Sigma),\mu_i)\}\right). \tag{EC3}$$

It is easy to show the existence of a unique equilibrium λ that satisfies (EC1)–(EC3). Essentially, condition (EC1) specifies that, in equilibrium, any provider i with higher total cost than some provider j will have no customers. Similarly, condition (EC2) specifies that, in equilibrium, any two providers with positive arrivals will have the same total costs. Together, these two conditions ensure that no customer can increase their utility by switching from the more expensive to the cheaper provider and imply (EC3).

The objective function for the providers is still given by (4) with $\lambda_i(\Sigma)$ replacing $\lambda(p,\mu)$ for provider i. Similarly, the objective of the regulator can be written as:

$$S(\Sigma)=V(\Xi(\Sigma))-\sum_{i=1}^N\left(tW(\lambda_i(\Sigma),\mu_i)+c_i\right)\lambda_i(\Sigma)-\sum_{i=1}^N R(c_i,\mu_i), \tag{EC4}$$

subject to (9), where $V(y)=\Lambda\int_{\bar{\Theta}^{-1}(y/\Lambda)}^{\infty} x d\Theta(x)$.

Throughout this section, we assume that there is a symmetric optimal solution for the regulator's optimization problem, denoted by $\hat{\Sigma}^*=(\hat{c}^*,\hat{\mu}^*,\hat{p}^*)$. This assumption implies that the regulator finds it optimal to have all providers active. We also assume that FOCs are necessary and sufficient for optimality.

We begin the analysis with the case when customers may be charged a fee for service. We first establish that regulatory intervention is still needed despite competition.

Proposition EC1. *If providers compete and are free to choose their own prices, capacity, and cost-reduction effort, first-best outcomes cannot be achieved in any equilibrium.*

This result follows from the observation that, if all firms were to choose first-best actions, then at least one firm will find it optimal to deviate. Therefore, first-best actions cannot be an equilibrium outcome. A consequence of the proposition above is that direct competition between firms is not sufficient to incentivize optimal behaviour in equilibrium. We next show that the cost-based yardstick competition of Proposition 2 is not effective in incentivizing optimal capacity investment.

Proposition EC2. *If providers compete and customers experience costly waiting time ($t > 0$), under the cost-based yardstick competition of Proposition 2, where the price, p_i, and transfer payment, T_i, are set by the regulator as given in (14) and (15), the firms' installed capacity is the minimum capacity level, μ_o, in all symmetric equilibria.*

This result establishes that demand-side competition between providers is not sufficient to reinstate the optimality of the cost-based yardstick competition. Despite the fact that any investment in capacity will increase the demand of any individual provider at the expense of other providers, in equilibrium, no provider wants to add capacity as the marginal value of the additional demand is zero (because in equilibrium the reimbursement by the regulator is equal to providers' cost). We next show that the yardstick regulation that was shown to restore first-best outcomes in the monopoly setting (Theorem 1), leads to first-best outcomes when providers compete as well.

Proposition EC3. *Under the cost- and capacity-based yardstick competition scheme of Theorem 1, where price, p_i, and transfer payment, T_i, are set by the regulator as given in* (16) *and* (17) *(with $\lambda_i(\Sigma)$ replacing the arrival rate λ_i for provider i in the definitions), the unique symmetric Nash equilibrium is for each provider i to pick the first-best outcomes $(\hat{c}^*, \hat{\mu}^*)$. Also, all providers make zero profit in equilibrium.*

We now turn to the case where customers are not charged a fee to access the service, corresponding to the analysis in Section 4.4 in the monopoly setting. We note that the objectives of the regulator and the providers, as well as the arrival rates, are identical to those given above except $p_i = 0$ for all $i = 1, 2, \ldots, N$. We again assume that a symmetric welfare maximizing solution exists for the regulator's problem, denoted by $\hat{\Sigma}^* = (\hat{c}_o^*, \hat{\mu}_o^*)$ (with a slight abuse of notation). We refer to this solution as the second-best. Unlike the case of first-best regulation, which was shown to be robust to competition, we next show that the payment scheme proposed for the monopoly case fails to generate second-best outcomes when firms compete.

Proposition EC4. *If customers are not charged directly (i.e., $p_i = 0$), and the regulator makes a transfer payment, T_i, defined as in* (23) *to provider i, for $i = 1, \ldots, N$, the firms' installed capacity is the minimum capacity level μ_o in all symmetric equilibria.*

The proposed payment scheme fails to incentivize any capacity investment simply because competition renders the benchmarking of waiting times irrelevant. Recall that, under this scheme, each service provider is paid a fee which is proportional to the difference between the average waiting time of every other provider and their own average waiting time. But if providers compete, in equilibrium, all (active) providers will have the same waiting time, irrespective of their individual capacity. Therefore, any investment in increasing capacity by provider i will be costly without generating any reduction in waiting time for provider i compared to other providers.

To circumvent this issue, while still using a payment scheme that is relatively easy to implement, we propose to set the wait-time benchmark using a similar market that operates independently from the one we consider. To demonstrate, consider a case with $2N$ identical service providers, where providers 1 through N, serve a different market to providers $(N+1)$ through $2N$, for example, EDs in different urban areas. For each provider i, we define sets $\mathbb{S}_i$ and $\mathbb{D}_i$ as the indices of providers who operate in the same area as provider i, and of those in the other area, respectively, for example, $\mathbb{S}_1 = \{1,...,N\}$ and $\mathbb{D}_1 = \{N+1,...,2N\}$. Consider the reimbursement scheme with provider i's transfer payment equal to

$$T_i = t\left(\bar{W}_i^D - \bar{W}_i^S\right) N\lambda_i + \bar{R}_i + \left(\sum_{j\in\mathbb{D}_i} c_j\lambda_j - \sum_{j\in\mathbb{S}_i, j\neq i} c_j\lambda_j\right), \tag{EC5}$$

where $\bar{W}_i^D$, $\bar{W}_i^S$, and $\bar{R}_i$ are given by

$$\bar{W}_i^D = \frac{1}{|\mathbb{D}_i|}\sum_{j\in\mathbb{D}_i} W_j, \quad \bar{W}_i^S = \frac{1}{|\mathbb{S}_i|}\sum_{j\in\mathbb{S}_i} W_j, \text{ and } \bar{R}_i = \frac{1}{|\mathbb{D}_i|}\sum_{j\in\mathbb{D}_i} R_j, \tag{EC6}$$

and we use λ_i and W_i to denote the (equilibrium) arrival rate and the expected waiting time for provider i. There are two main differences between this payment scheme and the scheme proposed for the monopoly case, see (23): i) The waiting-time benchmark and the lump-sum compensation $(\bar{R}_i)$ are based on a set of providers that do not compete with provider i, and ii) the reward for reducing wait time for each provider is multiplied by the number of competing providers, N. The first difference helps to reinstate the power of relative benchmarking which competition had eroded. The second difference serves to amplify the reward for providers who invest in reducing wait times, which due to competition, is not localized but marketwide.

Proposition EC5. *If the regulator makes the transfer payment, T_i, given in* (EC5) *to provider i, for $i = 1,...,2N$, and customers are not charged directly, the unique symmetric Nash equilibrium is for each provider to choose the second-best outcomes $(\hat{c}_o^*, \hat{\mu}_o^*)$. Also, all providers make zero profit in equilibrium.*

In summary, the presence of competition does not affect the first-best yardstick competition in any way. The second-best scheme is affected but this can be circumvented if, for every provider the regulator can identify a set of non-competing and a set of competing providers, and set the waiting-time and cost benchmarks based on the first set and the rewards based on the size of the second set.

EC.1.1. Proofs of the Results in Section EC.1

We first present some properties of equilibrium points that we later use in the proofs. Throughout we use "d" for derivative and "∂" for partial derivative.

LEMMA EC1. *Let $\Sigma=(c,\mu,p)$ be such that $\lambda_i(\Sigma)>0$ for all $i=1,2,\ldots,N$. Then*

i) $\lambda_j(\cdot)$ is increasing in μ_j and decreasing in p_j and $\lambda_i(\cdot)$ is decreasing in μ_j and increasing in p_j, for all $i,j=1,2,\ldots,N$,

ii) $\lambda_i(\cdot)$ is differentiable with respect to μ_i and p_i for all $i,j=1,2,\ldots,N$, at Σ and

$$\frac{dW\left(\lambda_i(\Sigma),\mu_i\right)}{d\mu_i}=\frac{dW\left(\lambda_j(\Sigma),\mu_j\right)}{d\mu_i},\quad i,j=1,2,\ldots,N. \tag{EC7}$$

The proof follows from (EC2)–(EC3) and the implicit function theorem using standard arguments, hence we omit the details.

Proof of Proposition EC1: Let Σ_i and Σ_{-i} denote the actions of provider i and of all providers except provider i, and let $\hat{\Sigma}_i^*$ and $\hat{\Sigma}_{-i}^*$ denote first-best actions (under competition) for provider i and of all providers, except provider i, respectively, for $i=1,...,N$.

We first show that the first-best price, $\hat{p}^*$, cost per customer, $\hat{c}^*$, capacity, $\hat{\mu}^*$, and transfer payment, $\hat{T}^*$, (in the competition case) are the unique solution to (10)–(13). We set $\hat{\Sigma}^*=(\hat{c}^*,...,\hat{c}^*,\hat{\mu}^*,...,\hat{\mu}^*,\hat{p}^*,...,\hat{p}^*)$. By (EC4), $\hat{c}^*$, $\hat{\mu}^*$, and $\lambda_i(\hat{\Sigma}^*)$ satisfy (10) because the optimal solution is symmetric.

By Leibniz rule

$$\frac{\partial}{\partial y}\left(\Lambda\int_{\bar{\Theta}^{-1}\left(\frac{y}{\Lambda}\right)}^{\infty}xd\Theta(x)\right)=-\Lambda\bar{\Theta}^{-1}\left(\frac{y}{\Lambda}\right)\theta\left(\bar{\Theta}^{-1}\left(\frac{y}{\Lambda}\right)\right)\frac{\partial\bar{\Theta}^{-1}\left(\frac{y}{\Lambda}\right)}{\partial y}=\bar{\Theta}^{-1}\left(\frac{y}{\Lambda}\right)\ \text{for } y\in[0,\Lambda]. \tag{EC8}$$

Let $v:[0,\Lambda]\to[0,\infty)$ be defined by

$$v(\lambda)=\bar{\Theta}^{-1}(\lambda/\Lambda). \tag{EC9}$$

In words, $v(\lambda)$ is the marginal value of the marginal customer (i.e., the customer who is indifferent between joining a provider to receive service and not joining) when the arrival rate is λ.

By (EC4) and (EC8)

$$\frac{\partial S(\Sigma)}{\partial p_j}=\frac{\partial\Xi(\Sigma)}{\partial p_j}v\left(\Xi(\Sigma)\right)-\sum_{i=1}^{N}\left(t\frac{\partial\lambda_i(\Sigma)}{\partial p_j}\frac{\partial W\left(\lambda_i(\Sigma),\mu_i\right)}{\partial\lambda}\right)\lambda_i\left(\Sigma\right)$$

$$
\begin{aligned}
&-\sum_{i=1}^{N}\left(tW\left(\lambda_i(\Sigma),\mu_i\right)+c_i\right)\frac{\partial\lambda_i\left(\Sigma\right)}{\partial p_j}\\
=&\frac{\partial\Xi\left(\Sigma\right)}{\partial p_j}\left(p_i-c_i-t\lambda_i(\Sigma)\frac{\partial W\left(\lambda_i(\Sigma),\mu_i\right)}{\partial\lambda}\right),
\end{aligned}
$$

where the second equality follows from our assumption that the optimal solution is symmetric. Hence (12) holds in this case as well.

Now

$$
\begin{aligned}
\frac{\partial S(\Sigma)}{\partial\mu_j}=&\frac{\partial\Xi(\Sigma)}{\partial\mu_j}v\left(\Xi(\Sigma)\right)-\sum_{i=1}^{N}\left(t\frac{\partial\lambda_i(\Sigma)}{\partial\mu_j}\frac{\partial W\left(\lambda_i(\Sigma),\mu_i\right)}{\partial\lambda}\right)\lambda_i\left(\Sigma\right)\\
&-t\frac{\partial}{\partial\mu}W\left(\lambda_j(\Sigma),\mu_j\right)\lambda_j\left(\Sigma\right)-\sum_{i=1}^{N}\left(tW\left(\lambda_i(\Sigma),\mu_i\right)+c_i\right)\frac{\partial\lambda_i\left(\Sigma\right)}{\partial\mu_j}-\frac{\partial}{\partial\mu}R(c_j,\mu_j)\\
=&-t\frac{\partial}{\partial\mu}W\left(\lambda_j(\Sigma),\mu_j\right)\lambda_j\left(\Sigma\right)-\frac{\partial}{\partial\mu}R(c_j,\mu_j),
\end{aligned}
$$

where the first equality follows from (EC8), and the second equality follows again from our assumption that the optimal solution is symmetric and from (12). Hence (11) holds in this case as well. The first-best transfer payment, $\hat{T}^*$, is obtained by solving for $\Pi(\hat{c}^*,\hat{\mu}^*|\hat{p}^*,\hat{T}^*)=0$, which leads to (13) by (4).

We next show that, if all providers $j\neq i$ choose first-best outcomes $\hat{\Sigma}^*_{-i}$, provider i's best response is not $\hat{\Sigma}^*_i$. By (4)

$$
\Pi_i(c_i,\mu_i|p_i,T_i,\Sigma_{-i})=(p_i-c_i)\lambda_i(\Sigma)-R(c_i,\mu_i)+T_i,\qquad i=1,...,N. \tag{EC10}
$$

By (11), (12), (EC10), and the fact that we assume the optimal solution is symmetric we have

$$
\begin{aligned}
\frac{\partial}{\partial\mu_i}\Pi_i(\hat{p}^*,\hat{c}^*,\hat{\mu}^*|\hat{T}^*_i,\hat{\Sigma}^*_{-i})&=(\hat{p}^*-\hat{c}^*)\frac{\partial\lambda_i(\hat{\Sigma}^*)}{\partial\mu_i}-\frac{\partial R(\hat{c}^*,\hat{\mu}^*)}{\partial\mu}\\
&=t\lambda_i(\hat{\Sigma}^*)\frac{\partial}{\partial\lambda}W(\lambda_i(\hat{\Sigma}^*),\hat{\mu}^*)\frac{\partial\lambda_i(\hat{\Sigma}^*)}{\partial\mu_i}-\frac{\partial R(\hat{c}^*,\hat{\mu}^*)}{\partial\mu}\\
&=t\lambda_i(\hat{\Sigma}^*)\frac{dW_i(\lambda_i(\hat{\Sigma}^*),\hat{\mu}^*)}{d\mu_i}.
\end{aligned}
\tag{EC11}
$$

Hence, it is enough to show that $\frac{dW_i(\lambda_i(\hat{\Sigma}^*),\hat{\mu}^*)}{d\mu_i}\neq 0$ to complete the proof. Because θ (the derivative of Θ) is non-zero by assumption, v is differentiable and $v'(\lambda)<0$ for all $\lambda\geq 0$. Hence there exists $\epsilon>0$ such that

$$
v'(\Xi(\hat{\Sigma}^*))<-\epsilon. \tag{EC12}
$$

Now assume that $\frac{dW_i(\lambda_i(\hat{\Sigma}^*),\hat{\mu}^*)}{d\mu_i}=0$. By Lemma EC1(ii) this implies that $\frac{dW_j(\lambda_j(\hat{\Sigma}^*),\hat{\mu}^*_j)}{d\mu_i}=0$ for all $j=1,2,\ldots,N$. Hence, there exists a sequence $\delta(n)\downarrow 0$ such that for $\mu_i(n)=\hat{\mu}_i+\delta(n)$

$$
W(\lambda_i(\hat{\Sigma}^*(n)),\mu_i(n))\geq W(\hat{\lambda}_i(\hat{\Sigma}^*),\hat{\mu}^*_i)-\delta_n\epsilon/2,\ \text{and} \tag{EC13}
$$

$$W(\lambda_j(\hat{\Sigma}^*(n)),\hat{\mu}_j^*) \geq W(\lambda_j(\hat{\Sigma}^*),\hat{\mu}_j^*) - \delta_n\epsilon/2,\ j\in\{1,\ldots,N\}\setminus i, \tag{EC14}$$

where $\hat{\Sigma}^*(n)$ is equal to $\hat{\Sigma}^*$ except for service rate at server i, which is equal to $\mu_i(n)$. By (EC3), (EC13), and (EC14)

$$v(\Xi(\hat{\Sigma}_n^*)) > v(\Xi(\hat{\Sigma}^*)) - \delta_n\epsilon/2. \tag{EC15}$$

This contradicts (EC12). □

Proof of Proposition EC2: The proof is identical to that of Proposition 3 and is omitted for brevity.

Proof of Proposition EC3: We show that the cost- and capacity-based yardstick payment scheme achieves first-best in equilibrium using the fact that the first-best price, $\hat{p}^*$, cost per customer, $\hat{c}^*$, capacity, $\hat{\mu}^*$, and transfer payment, $\hat{T}^*$, are the unique solution to (10)–(13) – a result we proved while proving Proposition EC1 – under the assumption that FOCs are necessary and sufficient to obtain the optimal actions of each provider. By (17), provider i's objective function is

$$\Pi_i(c_i,\mu_i|p_i,T_i,\Sigma_{-i}) = (\bar{c}_i - c_i)\bar{\lambda}_i(\Sigma) - R(c_i,\mu_i) + \bar{R}_i + t\bar{\lambda}_i(\Sigma)\frac{\partial}{\partial\mu}W(\bar{\lambda}_i(\Sigma),\bar{\mu}_i)(\bar{\mu}_i - \mu_i). \tag{EC16}$$

Let $a_j = (\tilde{c},\tilde{\mu})$ denote the actions of provider j for all $j\neq i$. Then, by (EC16), the FOCs of provider i's profit function are

$$\begin{aligned}\frac{\partial}{\partial\mu_i}\Pi_i(c_i,\mu_i|p_i,T_i,\tilde{\Sigma}_{-i}) &= (\tilde{c}-c_i)\frac{\partial\bar{\lambda}_i(\Sigma)}{\partial\mu_i} - t\bar{\lambda}_i(\Sigma)\frac{\partial}{\partial\mu}W(\bar{\lambda}_i(\Sigma),\tilde{\mu}) \\ &\quad + \frac{d}{d\mu_i}\left(t\bar{\lambda}_i(\Sigma)\frac{\partial}{\partial\mu}W(\bar{\lambda}_i(\Sigma),\tilde{\mu})\right)(\tilde{\mu}-\mu_i) - \frac{\partial}{\partial\mu_i}R(c_i,\mu_i),\end{aligned} \tag{EC17}$$

$$\begin{aligned}\frac{\partial}{\partial c_i}\Pi_i(c_i,\mu_i|p_i,T_i,\tilde{\Sigma}_j) &= -\frac{\partial}{\partial c_i}R(c_i,\mu_i) - \bar{\lambda}_i(\Sigma) + (\tilde{c}-c_i)\frac{\partial\bar{\lambda}_i(\Sigma)}{\partial c_i} \\ &\quad + \frac{d}{dc_i}\left(t\bar{\lambda}_i(\Sigma)\frac{\partial}{\partial\mu}W(\bar{\lambda}_i(\Sigma),\tilde{\mu})\right)(\tilde{\mu}-\mu_i),\end{aligned} \tag{EC18}$$

where $\tilde{\Sigma}_{-i}$ is the vector of cost, $\tilde{c}$, capacity, $\tilde{\mu}$, and price, p_j, for all firms $j\neq i$.

If $c_j = \hat{c}^*$ and $\mu_j = \hat{\mu}^*$ for all $j = 1,...,N$, then $\frac{\partial}{\partial\mu_i}\Pi_i = 0$ and $\frac{\partial}{\partial c_i}\Pi_i = 0$ by (10)–(12). Also because (10)–(12) have a unique solution, so do (16), (EC17), and (EC18). In addition, because FOCs are necessary and sufficient to obtain the optimal actions of each provider, $(\hat{c}^*,\hat{\mu}^*)$ is a Nash equilibrium where, clearly, providers make zero profit by (EC16). The uniqueness of the symmetric Nash equilibrium follows from the uniqueness of the outcomes that satisfy the FOCs of the welfare function, S, similar to the argument in the proof of Theorem 1. □

Proof of Proposition EC4: Let $\Gamma = (\mu_1,...,\mu_N)$ denote the vector of capacities of all providers and similarly let Γ_{-i} denote the capacities of all providers except provider i. Let $\Pi_i(c_i,\mu_i|T_i,\Gamma_{-i})$ denote the profit of provider i given by (24) with $\lambda(\Gamma)$ replacing $\lambda(\mu)$ in all relevant definitions. By assumption $\frac{\partial R(c_i,\mu_i)}{\partial\mu_i} > 0$ and by Lemma EC1, $\frac{\partial}{\partial\mu_i}\Pi_i(c_i,\mu_i|T_i,\Gamma_{-i}) < 0$ for all $\mu_i \geq \mu_o$ and c_i, giving the desired result. □

Proof of Proposition EC5: Let $\hat{\mu}_o^*$ and $\hat{c}_o^*$ denote optimal capacity and cost values for the regulator when the customers are not charged directly for service in the competition case and $\hat{\Gamma}^* = (\hat{\mu}_o^*, ..., \hat{\mu}_o^*)$. We denote the equilibrium arrival rate to provider i by $\lambda_i(\Gamma)$ when customers are not charged (i.e. $p_i = 0$ for all i) and providers choose the capacity vector Γ. Similar to the proof of Proposition EC3, by (EC4) the FOCs for welfare-maximizing actions with no fee are given by

$$\frac{\partial R(\hat{c}_o^*, \hat{\mu}_o^*)}{\partial \mu_i} = -\sum_{j=1}^{N} \left(t\lambda_j(\hat{\Gamma}^*) \frac{dW_j(\lambda_j(\hat{\Gamma}^*), \hat{\mu}_o^*)}{d\mu_i} + \hat{c}_o^* \frac{\partial \lambda_j(\hat{\Gamma}^*)}{\partial \mu_i} \right), \; i = 1, ..., N, \tag{EC19}$$

$$\frac{\partial R(\hat{c}_o^*, \hat{\mu}_o^*)}{\partial c} = -\lambda_i(\hat{\Gamma}^*), \tag{EC20}$$

$$\hat{T}_o^* = \hat{c}_o^* \lambda_i(\hat{\Gamma}^*) + R(\hat{c}_o^*, \hat{\mu}_o^*). \tag{EC21}$$

Recall that we consider the special case with providers 1 through N, and providers $(N+1)$ through $2N$ serve a different market. Therefore, the arrival rate of provider i depends only on the capacity of providers that operate in the same market, that is,

$$\frac{\partial \lambda_j(\hat{\Gamma}^*)}{\partial \mu_i} = 0 \text{ for } j \in \mathbb{D}_i. \tag{EC22}$$

By (EC5), provider i's profit under this payment scheme is given by

$$\Pi_i(c_i, \mu_i | T_i, \Sigma_{-i}) = -\sum_{j \in \mathbb{S}_i} c_j \lambda_j(\Gamma) + t\left(\bar{W}_i^D - \bar{W}_i^S\right) N\lambda_i(\Gamma) - R_i(c_i, \mu_i) + \bar{R}_i + \sum_{j \in \mathbb{D}_i} c_j \lambda_j(\Gamma), \quad i = 1, ..., 2N, \tag{EC23}$$

where $\bar{W}_i^D$, $\bar{W}_i^S$, and $\bar{R}_i$ are as given in (EC6). We next show that there is a unique symmetric equilibrium. Let $a_j = (\tilde{c}, \tilde{\mu})$ denote the actions of provider j for all $j \neq i$. Without loss of generality, we focus on the profit of provider 1, for which $\mathbb{S}_1 = \{1, ..., N\}$ and $\mathbb{D}_1 = \{N+1, ..., 2N\}$. By (EC22) and (EC23), the FOCs for provider 1's profit function are

$$\frac{\partial}{\partial \mu_1} \Pi_1(c_1, \mu_1 | T_1, \Sigma_{-1}) = -\sum_{j=1}^{N} c_j \frac{\partial \lambda_j(\Gamma)}{\partial \mu_1} - t\lambda_1(\Gamma) \sum_{j=1}^{N} \frac{d}{d\mu_1} W(\lambda_j(\Gamma), \mu_j) - \frac{\partial R(c_1, \mu_1)}{\partial \mu_1}, \tag{EC24}$$

$$\frac{\partial}{\partial c_1} \Pi_1(c_1, \mu_1 | T_1, \Sigma_{-1}) = -\lambda_1(\Gamma) - \frac{\partial R(c_1, \mu_1)}{\partial c_1}. \tag{EC25}$$

If $\tilde{c} = \hat{c}_o^*$ and $\tilde{\mu} = \hat{\mu}_o^*$, then by (EC19), (EC20), (EC24) and (EC25), $c_1 = \hat{c}_o^*$ and $\mu_1 = \hat{\mu}_o^*$ would yield $\frac{\partial \Pi_1}{\partial \mu_1} = 0$ and $\frac{\partial \Pi_1}{\partial c_1} = 0$. Because FOCs are assumed to be necessary and sufficient to obtain providers' optimal actions, $(\hat{c}_o^*, \hat{\mu}_o^*)$ is a Nash equilibrium. Because (EC19) and (EC20) have a unique solution, so do (EC24) and (EC25); and thus, $(\hat{c}_o^*, \hat{\mu}_o^*)$ is the unique symmetric Nash equilibrium. Finally, by (EC23) and because $|\mathbb{S}_i| = |\mathbb{D}_i|$, providers' equilibrium profit is zero. □

EC.2. Modeling Extensions

In this section, we discuss how yardstick competition could be implemented under more general conditions than those of §4. Namely, we look at the case of multiple customer classes, time-varying arrival rates, more general cost structure, using tail-statistics instead of average waiting time, provider heterogeneity, purely exogenous arrivals, and more general queueing models. Although it is possible to present these extensions for the first-best yardstick competition of §4.3, we have chosen to restrict attention to the simpler second-best yardstick competition of §4.4, which we believe to be of more practical relevance. For all extensions, we assume that the FOCs are necessary and sufficient for determining the unique solution to the regulator's welfare maximization problem and that each provider's objective is concave with the new payment scheme for each extension.

EC.2.1. Multiple Customer Classes and Time-varying Arrival Rates

In most practical settings, there are multiple customer classes with different service needs utilizing the same limited resources, and there are settings where the arrival rates are time varying. For example, in emergency care patients are triaged into different levels based on their severity, and arrival rates are much higher during the day than the evenings (Armony et al. 2015). The way that the payment mechanisms need to be modified to account for these two additional features are somewhat similar, so we focus on the extension for multiple customer classes and then explain how the time-varying arrivals can be handled in a similar manner.

Assume for simplicity that each provider caters to two different customer classes. We use the model presented in §3.1 for customer joining behavior, but we append a superscript j to denote the quantity associated with each customer class j, that is, $\Lambda^{(j)}, \Theta^{(j)}$, and $t^{(j)}$ are all assumed to depend on the customer class. Because customers from both classes use the same resources, the average waiting time of each class not only depends on the service rate and the arrival rate for that class but also depends on those of the other class, as well as the priority policy. We assume that all providers follow the same priority policy and that the class a customer belongs to is observable to the provider. For example, in emergency care patients are prioritized according to their severity levels and a similar triage method is used across hospital EDs (see McHugh et al. (2012) and Gilboy et al. (2011)).

The service rate of type i customers is denoted by $\mu^{(i)}$ and we assume that the provider can invest in increasing service rates and/or reducing cost per patient at a cost given by $R(c, \mu^{(1)}, \mu^{(2)})$. Let $W^{(j)}(\mu^{(1)}, \mu^{(2)}, \lambda^{(1)}, \lambda^{(2)})$ denote the expected waiting time for customer class j if the service and arrival rates of each class are given by $\mu^{(j)}$ and $\lambda^{(j)}$, $j = 1, 2$, where $\lambda^{(j)}$ satisfies (2) (with the average waiting time definition extended as described). For notational simplicity, we denote the expected wait by $W^{(j)}$ and in provider i by $W_i^{(j)}$ for class j. The objective of the regulator with two customer classes is

$$S\left(c,\mu^{(1)},\mu^{(2)}\right)=\sum_{j=1}^{2}\Lambda^{(j)}\int_{t^{(j)}W^{(j)}}^{M_r}\left(x-t^{(j)}W^{(j)}\right)d\Theta^{(j)}(x)-c\left(\lambda^{(1)}+\lambda^{(2)}\right)-R\left(c,\mu^{(1)},\mu^{(2)}\right). \tag{EC26}$$

Let c^*, $\mu^{(1),*}$ and $\mu^{(2),*}$ denote the optimal solution to (EC26). The model can easily be extended to the case for which the cost c depends on the customer class as well. Set the transfer payment to provider i to

$$T_i=t^{(1)}\left(\bar{W}_i^{(1)}-W_i^{(1)}\right)\bar{\lambda}_i^{(1)}+t^{(2)}\left(\bar{W}_i^{(2)}-W_i^{(2)}\right)\bar{\lambda}_i^{(2)}+\bar{R}_i+\bar{c}_i\left(\bar{\lambda}_i^{(1)}+\bar{\lambda}_i^{(2)}\right), \tag{EC27}$$

where, similar to (15),

$$\bar{\lambda}_i^{(j)}=\frac{\sum_{k\neq i}\lambda_k^{(j)}}{N-1},\ \bar{W}_i^{(j)}=\frac{\sum_{k\neq i}W_k^{(j)}}{N-1}\text{ and }\bar{R}_i=\frac{\sum_{k\neq i}R\left(c_k,\mu_k^{(1)},\mu_k^{(2)}\right)}{N-1},\ i=1,\dots,N,\text{ and }j=1,2. \tag{EC28}$$

Then, in the unique symmetric equilibrium, each provider picks c^*, $\mu^{(1),*}$, and $\mu^{(2),*}$. The proof is very similar to that of Proposition 2 and, in the interest of brevity, is omitted. We note that the transfer payment in the case of multiple patient classes given in (EC27) is a rather straightforward extension to that of the single patient class. It consists of the *total* expected cost $\left(\bar{R}_i+\bar{c}_i\left(\bar{\lambda}_i^{(1)}+\bar{\lambda}_i^{(2)}\right)\right)$ of providing service across all customers (i.e., it is not necessary to know the exact breakdown of this cost across different customer classes), plus a customer-class-specific fee that is increasing in the difference between the industry-wide and the specific provider waiting time for that customer class. Extensions to more customer classes are also straightforward.

If arrival rates are time-varying, the proposed scheme can be modified similarly if the potential and actual arrival rates (Λ and λ in our notation, respectively) are assumed to remain constant in non-overlapping intervals, and if expected waiting times for any customers arriving in the same interval are the same. Although this approach ignores the transient behavior of the queues going from one interval to the other, if the service times are relatively short compared to the rate of change in arrivals, it should yield accurate results. This approach is widely used in call centers, see Gans et al. (2003). If service times are longer than the rate of change in arrivals, as may be the case in EDs, then two customers may experience different waiting times, even though their times of arrival are not far apart. One potential solution in this case is to divide the day into multiple smaller intervals. For example, Armony et al. (2015) has documented that there are four distinct intervals in terms of occupancy in an Israeli ED with two relatively “stationary” periods, first from 3am to 9am and second from 12pm to 11pm and two transient periods in between these two stationary periods. Theoretically, by dividing the transient periods into multiple periods, one can ensure that waiting times experienced by patients who arrived in these sub-intervals are similar.

To demonstrate the yardstick scheme in a system with time-varying arrivals, assume that there are two intervals in one day during which the arrival rates are assumed to be constant and customers who arrive in the same interval experience the same expected waiting times. We consider only one customer class and assume that customers do not choose strategically when to arrive. With a slight abuse of notation let $W^{(j)}(\mu^{(1)}, \mu^{(2)}, \lambda^{(1)}, \lambda^{(2)})$ denote the expected waiting time for customers arriving during period j if the service and arrival rates during each time interval are given by $\mu^{(j)}$ and $\lambda^{(j)}$, $j = 1, 2$. Then, if the regulator sets the transfer payment, as in (EC27) (with the averages defined over time intervals instead of customer classes), it can be shown that socially optimal choices for the providers is a unique symmetric Nash equilibrium.

In summary, in the presence of multiple customer classes and time-varying arrival rates, the regulator needs to augment the yardstick competition model of §4.4 by making the transfer payment contingent on the relative performance of the provider for each customer class and each time interval. For example, in the case of ED, the day can be broken into four intervals (as per Armony et al. (2015)) and average wait times for each of the five triage levels (as per McHugh et al. (2012)) can be compared in these intervals.

EC.2.2. Extending the Cost Model

The model we used for the regulator's and the providers' objective can be extended to the case when the cost per customer also depends on the service rate, for example, treating patients faster affects the cost per patient. We could do this by assuming that the marginal cost, c, is no longer a decision variable itself, but is instead a non-negative valued function, $c(e, \mu)$, of costly effort (denoted by e), and the service rate μ. Similarly, the total investment cost would also be the function $R(e, \mu)$. The mechanisms proposed in §§4.3–4.4 would still result in first- and second-best outcomes in equilibrium with the definitions of $\bar{c}_i$ and $\bar{R}_i$ modified as follows

$$\bar{c}_i = \frac{1}{N-1} \sum_{j \neq i} c_j(e, \mu) \text{ and } \bar{R}_i = \frac{1}{N-1} \sum_{j \neq i} R(e_j, \mu_j)$$

in payment schemes (16)-(17) and (23).

EC.2.3. Yardstick Competition Using Tail Statistics

An alternative to incentivizing service providers based on their average wait is to use the tail statistics of their wait time, for example, the fractile of the wait time distribution. For example, in order to incentivize EDs to reduce their wait times, Monitor, the UK hospital regulator, mandates EDs to admit or discharge 95% of the patients within four hours of their arrival, and financially penalizes the hospitals that fail to reach this target (Campbell 2016). Since the welfare-maximizing performance target cannot be computed by the regulator without knowing the cost function $R(c,\mu)$ or the patient equilibrium arrival rate $\lambda(c,\mu)$, the yardstick competition that we propose in §§4.3–4.4 can be modified to achieve first- and second-best outcomes by using tail statistics of wait time.

To demonstrate, assume that the utility of a customer is given by $r - t\Sigma(\lambda,\mu) - p$ for a non-negative function Σ (the model in §3.1 is a special case). For example, if the regulator believes that customers' utility depends on the probability of waiting more than four hours, then $\Sigma(\lambda,\mu) = \mathbb{E}[\mathbb{1}\{w(\lambda,\mu) \geq 4\}]$, where w is the (random) waiting time of a customer in steady state. Similarly, if the regulator believes that customers' utility depends on 95% fractile of waiting times, then $\Sigma(\lambda,\mu) = H^{-1}_{\lambda,\mu}(95\%)$, where $H_{\lambda,\mu}$ is the distribution of the waiting time in steady-state when arrival and service rates are λ and μ, respectively. Then, the proposed schemes still lead to first- and second-best outcomes in equilibrium if the function W is replaced by Σ in (16)–(17) or (22)–(24).

EC.2.4. Heterogeneous Hospitals

To implement the proposed regulatory schemes, it was assumed that the regulator was able to identify (at least pairs of) identical providers. In many real-world settings, such as hospital EDs, this might not be possible because hospitals may differ along multiple dimensions, for example, the size of their respective catchment areas Λ, the distribution of customer benefits $\Theta(.)$ (e.g., due to case mix variation), and due to differences in the local labor markets giving rise to different costs of treatment, c, capacity, μ, and investment cost, $R(c,\mu)$. Nevertheless, if the regulator is able to observe the characteristics that make the providers differ, the proposed schemes can be modified in a way similar to Shleifer (1985).

To illustrate, note that to implement the second-best yardstick competition the regulator needs to be able to project the total cost, total number of customer arrivals, and average waiting time of each provider. Assume that each provider exhibits a different total cost $f_i(\mu,c,\delta) = c_i(\delta)\lambda_i(\delta) + R_i(\delta)$, arrival rate $\lambda(\mu,\delta)$, and waiting time $W_i(\mu,\delta)$, where δ is a vector containing all observable characteristics that make providers different. Also, assume that δ is not under the control of the providers. Then the regulator can use the information on the total costs, arrival, and average waiting times of all other providers along with the vector of observable characteristics, to predict the expected costs of each provider. For example, this could be achieved through a multivariate panel regression or indeed any other method (e.g., machine learning). If the predictive model is 100% accurate in predicting costs, and all the observable characteristics are correctly accounted for, the proposed scheme generates the socially optimal equilibrium. Obviously, as the explanatory power of the model used by the regulator degrades, so will the value of using yardstick competition.[1]

[1] An alternative scheme, which can be particularly useful as the unexplained heterogeneity between providers is

EC.2.5. Exogenous Arrivals

In certain situations, patients might not have a choice but seek service from the provider regardless of wait times, for example, cancer patients seeking chemotherapy treatment. In those cases, λ is exogenous (i.e., not a function of μ) and can be assumed to be fixed in our model. Then it can be shown, as in Theorem 2, that the payment scheme (23) still yields the second-best outcomes in equilibrium. More specifically, the regulator's objective can be written in this case as

$$S(c,\mu) = V(\lambda) + (p-c)\lambda - R(c,\mu), \tag{EC29}$$

and the FOCs for an optimal point (c_o^*, μ_o^*) become

$$\begin{aligned}
&\frac{\partial}{\partial c} R(c_o^*, \mu_o^*) = -\lambda, \\
&\frac{\partial}{\partial \mu} R(c_o^*, \mu_o^*) = -t\lambda \frac{d}{d\mu} W(\lambda, \mu_o^*), \\
&T_o^* = R(c_o^*, \mu_o^*),
\end{aligned}$$

which are similar to those in Proposition 4.

Under the payment scheme defined in (23), the providers' objective becomes

$$\Pi(c_i, \mu_i | T_i) = -c\lambda + t(\bar{W}_i - W_i)\lambda - R(c_i, \mu_i) + \bar{R}_i + \bar{c}_i\lambda. \tag{EC30}$$

It is easy to show that the unique equilibrium is the second best by using the FOCs for the provider and the regulator (under the assumption that FOCs for the provider are necessary and sufficient). The proof is identical to that of Theorem 2, and hence is omitted.

EC.2.6. More Complex Queueing Systems

Motivated by analytical tractability, the model assumes that providers can only control waiting times through a single parameter μ. In general this may not be true, for example, in EDs, patients are treated using different resources in multiple treatment steps. Our results can be extended to cover such queueing networks as follows.

Assume that patients arrive at the provider i at rate λ_i and need to go through m stages that are served by n resources, for example, providers, bed or imaging equipment. Let $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in})$ denote the capacities of these n resources for provider i. Assume that the provider has control over $\boldsymbol{\mu}_i$ and the expected waiting times (W) are determined by λ_i and $\boldsymbol{\mu}_i$. Then it can be shown that the payment mechanism (23) still leads to second-best outcomes in equilibrium.

relatively high, is the modified yardstick competition proposed by Laffont and Tirole (1993), pp 84-86, where the regulator offers a menu of incentive-compatible yardstick-competition contracts and allows providers to self select. Although this may be more complex to implement, it has the potential to further reduce the inefficiency associated with asymmetric information and heterogeneous providers. It cannot, however, restore socially optimal outcomes and the more efficient providers retain positive rents.

To obtain the socially optimal outcomes in equilibrium in the first best, the payment scheme needs to be modified slightly to result in the following profit function:

$$\Pi(c_i,\mu_i|p_i,T_i)=(\bar{c}_i-c_i)\bar{\lambda}_i+t\bar{\lambda}_i\sum_{j=1}^{n}\frac{\partial}{\partial\mu_{ij}}W(\bar{\lambda}_i,\bar{\boldsymbol{\mu}}_i)(\bar{\mu}_{ij}-\mu_{ij})-R(c_i,\boldsymbol{\mu}_i)+\bar{R}_i, \tag{EC31}$$

where

$$\bar{\boldsymbol{\mu}}_i=\frac{1}{N-1}\sum_{j\neq i}\boldsymbol{\mu}_i \text{ and } \bar{\mu}_{ik}=\frac{1}{N-1}\sum_{j\neq i}\mu_{jk}. \tag{EC32}$$

We highlight the fact that the topology of the underlying queueing system of the ED is not important, for example, it could be a complex Jackson network, and our results still apply as long as the FOCs are necessary and sufficient for the regulator and the providers. However, increased complexity of the underlying queueing system imposes more informational burden on the regulator to implement the first-best scheme as the regulator needs to know the number of steps involved in treating a patient as well as the impact of the capacity of each step on waiting times ($\frac{\partial}{\partial\mu_{ij}}W(\bar{\lambda}_i,\bar{\boldsymbol{\mu}}_i)$), see (EC31). However, the required information to achieve second best is not affected.

EC.3. Conditions

As is often the case in queueing games, we have made the assumption that FOCs are necessary and sufficient for both the regulator's and the individual provider's problems (see for example Mendelson and Whang (1990)). This assumption, however, may not always hold (Stidham 2009). For this reason, in this section, we present additional conditions which ensure that FOCs are sufficient for the specific case of the $M/M/1$ queue. More specifically, we provide sufficient (but not necessary) conditions for:

- FOCs to be necessary and sufficient for determining the first- and second-best solutions of the regulator's welfare maximization problem of Propositions 1 and 4 (see Appendices EC.3.1 and EC.3.4, respectively).
- FOCs to be necessary and sufficient for the provider's profit maximization problem under the payment schemes of Propositions 2, Theorems 1 and 2 (see Appendices EC.3.2, EC.3.3, and EC.3.5, respectively).

Where applicable, we compare the conditions presented here to the closest extant literature (Shleifer 1985). We note that these conditions can all be interpreted as either assumptions about the cost function $R(c,\mu)$, which, loosely speaking, needs to be "sufficiently" convex for the objective functions to be well-behaved, or about the default capacity μ_o (and default cost c_o), which needs to be sufficiently small (large) to allow for interior solutions. We also note that these conditions are sufficient but by no means necessary.

EC.3.1. Assumptions for Proposition 1: Regulator's First-best Solution

We present sufficient conditions for the FOCs to be necessary and sufficient to obtain the regulators' optimal (first-best) actions. We begin by defining $\lambda^*(c,\mu) \in [0,\Lambda]$ by

$$\bar{\Theta}^{-1}\left(\frac{\lambda^*(c,\mu)}{\Lambda}\right) = \frac{t\mu}{(\mu-\lambda^*(c,\mu))^2} + c, \tag{EC33}$$

and $\mu^*(c) \geq \mu_o$ by

$$\frac{t\lambda^*(c,\mu^*(c))}{(\mu^*(c)-\lambda^*(c,\mu^*(c))^2} = \frac{\partial}{\partial\mu} R(c,\mu^*(c)) \tag{EC34}$$

for $c \in (0,c_o]$ and $\mu \geq \mu_o$. We show that $\lambda^*(c,\mu)$ and $\mu^*(c)$ are well-defined in the proof of Proposition EC6 below. We next define Conditions 1–5 that are used in the next proposition.

Condition 1: $\frac{\lambda^*(c,\mu)}{(\mu-\lambda^*(c,\mu))^2}$ is decreasing in μ for $\mu \geq \mu_o$ and $c \in (0,c_o]$,
Condition 2: $\frac{t\lambda^*(c,\mu_o)}{(\mu_o-\lambda^*(c,\mu_o))^2} - \frac{\partial R(c,\mu_o)}{\partial\mu} > 0$ for $c \in (0,c_o]$,
Condition 3: $\left(-\lambda^*(c,\mu^*(c)) - \frac{\partial R(c,\mu^*(c))}{\partial c}\right)$ is decreasing in c for $c \in (0,c_o]$,
Condition 4: $\lambda^*(0,\mu^*(0)) < -\frac{\partial R(0,\mu^*(0))}{\partial c}$,
Condition 5: $\lambda^*(c_o,\mu^*(c_o)) > -\frac{\partial R(c_o,\mu^*(c_o))}{\partial c}$.

PROPOSITION EC6. *Conditions 1–5 are sufficient for the FOCs to be necessary and sufficient for the regulator's welfare maximization problem to define the first-best solution.*

Conditions 2, 4, and 5 are the boundary conditions that guarantee that first-best outcomes are interior. Condition 1 ensures the concavity of welfare function in capacity μ for all $c \in (0,c_o]$ when customers join at rate $\lambda^*(c,\mu)$. For instance, Condition 1 holds if the utilization in the system is decreasing in capacity (for all cost levels) when customers join in a socially optimal manner, that is, $\frac{\lambda^*(c,\mu)}{\mu}$ is decreasing in μ. Condition 3 is a necessary and sufficient condition for the concavity of welfare function in cost per customer, c, when customers join at rate $\lambda^*(c,\mu)$ and the provider chooses capacity level $\mu^*(c)$.

The boundary conditions given in Conditions 2, 4, and 5 essentially impose constraints on the marginal cost of waiting-time- and cost-reduction investments. For instance, Conditions 2 and 5 require that the marginal costs of capacity and cost investments are sufficiently small at the initial capacity, μ_o, and cost, c_o. Alternatively, given the convexity of investment cost function, R, these conditions can also be interpreted as the initial capacity, μ_o, being sufficiently small and the initial cost, c_o, being sufficiently high. They imply that the regulator always prefers to invest in cost reduction or in capacity, even when the other parameters are not chosen optimally. Similarly, Condition 4 requires the marginal cost at zero cost per customer to be sufficiently large, for example, Condition 4 holds if marginal cost at $c = 0$ is greater than the size of the catchment area, Λ.

Conditions 3–5 are similar to the conditions in Shleifer (1985) (given in (EC45) and (EC46)) that guarantee that FOCs are sufficient if waiting time is not costly. The only difference is that Shleifer (1985) inherently assumes that capacity is fixed at its default level, μ_o, and hence defines conditions for capacity level μ_o instead of $\mu^*(c)$.

Proof of Proposition EC6: We proceed in three main steps.

Step i) First we show that for fixed $c \in [0, c_o]$ and $\mu \in [\mu_o, \infty)$ there exists a unique $p \geq 0$, denoted by $p^*(c,\mu)$, that maximizes $S(p,c,\mu)$, that is,

$$S(p^*(c,\mu),c,\mu) = \max_{p\geq 0} S(p,c,\mu).$$

Step ii) Then we show that for fixed $c \in [0, c_o]$, $\mu^*(c)$ defined in (EC34) is unique and maximizes $S(p^*(c,\mu),c,\mu)$, that is,

$$S(p^*(c,\mu^*(c)),c,\mu^*(c)) = \max_{\mu\geq\mu_o} S(p(c,\mu),c,\mu).$$

Step iii) Next we show that there exists a unique $c^* \in (0, c_o)$ that maximizes $S(p^*(c,\mu^*(c)),c,\mu^*(c))$, that is,

$$S(p^*(c,\mu^*(c^*)),c^*,\mu^*(c^*)) = \max_{0\leq c\leq c_o} S(p(c,\mu),c,\mu^*(c)).$$

We provide the details of these steps next.

Step i) To prove the existence and uniqueness of $p^*(c,\mu)$ we show that $S(p,c,\mu)$ is a concave function of p and $p^*(c,\mu)$ is not at the boundaries, that is, $p^*(c,\mu) \in (0,\infty)$, for fixed $c \in [0, c_o]$ and $\mu \in [\mu_o, \infty)$. By (7) we have

$$\frac{\partial}{\partial p} S(p,c,\mu) = f(p,c,\mu)\frac{\partial}{\partial p}\lambda(p,\mu), \tag{EC35}$$

where, by (EC8), f is given by

$$f(p,c,\mu) := \bar{\Theta}^{-1}\left(\frac{\lambda(p,\mu)}{\Lambda}\right) - \frac{t\mu}{(\mu-\lambda(p,\mu))^2} - c. \tag{EC36}$$

By (3) we have

$$\frac{\partial}{\partial p}\lambda(p,\mu) < 0, \text{ for } p \geq 0 \text{ and } \mu \geq \mu_o. \tag{EC37}$$

Therefore, to show $S(p,c,\mu)$ is concave in p it is enough to show that $f(p,c,\mu)$ is increasing in p. By (EC36)

$$\frac{\partial}{\partial p} f(p,c,\mu) = \left(v'(\lambda(p,\mu)) - \frac{2t\mu}{(\mu-\lambda(p,\mu))^3}\right)\frac{\partial}{\partial p}\lambda(p,\mu). \tag{EC38}$$

(We use $'$ to denote the derivative of functions whose domain is one dimensional). We have $v'(y) < 0$ for all $y \in [0,\Lambda]$ because $\theta(x) \geq 0$ for $x \geq 0$, and that $\lambda(p,\mu) < \mu$ by (3). Hence $f(p,c,\mu)$ is increasing in p.

To show that $p^*(c,\mu) > 0$ and is unique, we next show that $\frac{\partial}{\partial p} S(0,c,\mu) > 0$ and $\lim_{p\to\infty} \frac{\partial}{\partial p} S(p,c,\mu) < 0$. By (EC36)

$$f(0,c,\mu) = v(\lambda(0,\mu)) - \frac{t\mu}{(\mu - \lambda(0,\mu))^2} - c = \frac{t}{\mu - \lambda(0,\mu)} - \frac{t\mu}{(\mu-\lambda(0,\mu))^2} - c < 0, \tag{EC39}$$

where the second equality follows from (3), giving $\frac{\partial}{\partial p} S(0,c,\mu) > 0$. Also, since $\lim_{p\to\infty} \lambda(p,\mu) = 0$ for all $\mu \geq \mu_o$ by (3), we have

$$\lim_{p\to\infty} f(p,c,\mu) > 0, \tag{EC40}$$

giving $\lim_{p\to\infty} \frac{\partial}{\partial p} S(p,c,\mu) < 0$.

Therefore, for any $c \in [0,c_o]$ and $\mu \in [\mu_o,\infty)$, there exists a unique optimal $p^*(c,\mu) \in (0,\infty)$, which is obtained by

$$\frac{\partial}{\partial p} S(p^*(c,\mu),c,\mu) = v(\lambda(p^*(c,\mu),\mu)) - \frac{t\mu}{(\mu - \lambda(p^*(c,\mu),\mu))^2} - c = 0. \tag{EC41}$$

This also proves the existence and uniqueness of $\lambda^*(c,\mu)$ defined in (EC33).

Step ii) To prove that there exists a unique $\mu \in (\mu_o,\infty)$ that maximizes $S(p^*(c,\mu),c,\mu)$ for fixed $0 \leq c \leq c_o$, we show that (i) $S(p^*(c,\mu),c,\mu)$ is concave in μ. (ii) $\lim_{\mu\downarrow\mu_o} \frac{d}{d\mu} S(p^*(c,\mu),c,\mu) > 0$ and (iii) $\lim_{\mu\to\infty} \frac{d}{d\mu} S(p^*(c,\mu),c,\mu) < 0$. By (7) and (EC8),

$$\begin{aligned}\frac{d}{d\mu} S(p^*(c,\mu),c,\mu) &= \left(v(\lambda^*(c,\mu)) - \frac{t\mu}{(\mu-\lambda^*(c,\mu))^2} - c \right) \frac{\partial}{\partial\mu} \lambda^*(c,\mu) + \frac{t\lambda^*(c,\mu)}{(\mu-\lambda^*(c,\mu))^2} - \frac{\partial R(c,\mu)}{\partial\mu} \\ &= \frac{t\lambda^*(c,\mu)}{(\mu-\lambda^*(c,\mu))^2} - \frac{\partial R(c,\mu)}{\partial\mu}, \end{aligned} \tag{EC42}$$

where (EC42) follows from (EC41). By Condition 1 and that $\frac{\partial R(c,\mu)}{\partial\mu}$ is increasing in μ by convexity of R, $\frac{d}{d\mu} S(p^*(c,\mu),c,\mu)$ is decreasing; and hence, $S(p^*(c,\mu),c,\mu)$ is concave in μ. By Condition 2 and (EC42), we have $\frac{d}{d\mu} S(p^*(c,\mu_o),c,\mu_o) > 0$. Also, by (EC42), convexity of R and $\lambda^*(c,\mu) \leq \Lambda$ by (3), we get

$$\lim_{\mu\to\infty} \frac{d}{d\mu} S(p^*(c,\mu),c,\mu) = \lim_{\mu\to\infty} -\frac{\partial R(c,\mu)}{\partial\mu} < 0. \tag{EC43}$$

Thus, the optimal capacity $\mu^*(c)$ given any $c \in (0,c_o]$ can be obtained by (EC34).

Step iii) Finally, we show that there is a unique $c \in (0,c_o)$ that maximizes $S(p^*(c,\mu^*(c)),c,\mu^*(c))$. Let $S^*(c) = S(p^*(c,\mu^*(c)),c,\mu^*(c))$ for notational simplicity. By (7) and (EC8), we have

$$\frac{\partial S^*(c)}{\partial c} = \left(v(\lambda^*(c,\mu^*(c))) - \frac{t\mu^*(c)}{(\mu^*(c) - \lambda^*(c,\mu^*(c)))^2} - c \right) \frac{d}{dc} \lambda^*(c,\mu^*(c))$$

$$+\left(\frac{t\lambda^*(c,\mu^*(c))}{(\mu^*(c)-\lambda^*(c,\mu^*(c)))^2}-\frac{\partial R(c,\mu^*(c))}{\partial\mu}\right)\frac{\partial\mu^*(c)}{\partial c}-\lambda^*(c,\mu^*(c))-\frac{\partial R(c,\mu^*(c))}{\partial c}$$
$$=-\lambda^*(c,\mu^*(c))-\frac{\partial R(c,\mu^*(c))}{\partial c}, \tag{EC44}$$

where (EC44) follows from (EC33) and (EC34). By (EC44) and Condition 3, $\frac{dS^*(c)}{dc}$ is decreasing and hence $S^*(c)$ is strictly concave in c. In addition, by Conditions 4 and 5, we have $\frac{dS^*(0)}{dc}>0$ and $\frac{dS^*(c_o)}{dc}<0$. Thus, there exists a unique global maximum c^* of $S^*(c)$ and $(p^*(c^*,\mu^*(c^*)),\mu^*(c^*),c^*)$ is the unique optimizer of $S(c,p,\mu)$ and satisfies the FOCs (11)–(13). Because (EC41), (EC42) and (EC44) have unique solutions, $(p^*(c^*,\mu^*(c^*)),\mu^*(c^*),c^*)$ is the unique point that satisfies the FOCs. □

EC.3.2. Assumptions for Proposition 2

We provide sufficient conditions that guarantee the FOCs of the provider's profit maximization problem to be necessary and sufficient under the payment scheme of Proposition 2. Since this is the problem studied in Shleifer (1985), these conditions were first presented in Shleifer (1985). Expressed in our notation, these are that $R(c,\mu_o)$ is convex in c and

$$\frac{\partial\lambda(c,\mu_o)}{\partial c}+\frac{\partial^2 R(c,\mu_o)}{\partial c^2}>0, \tag{EC45}$$

along with the boundary conditions

$$\lambda(c_o,\mu_o)+\frac{\partial}{\partial c}R(c_o,\mu_o)>0 \text{ and } \lambda(0,\mu_o)+\frac{\partial}{\partial c}R(0,\mu_o)<0. \tag{EC46}$$

Further to Shleifer (1985), we need to also assume that (EC45) holds for all $\mu\geq\mu_o$. These conditions ensure that the objective function of the provider is concave and the solution is interior for all $\mu\geq\mu_o$.

EC.3.3. Assumptions for Theorem 1

We provide sufficient conditions that guarantee the FOCs of the provider's profit maximization problem to be necessary and sufficient under the payment scheme of Theorem 1. The provider's profit is given by (18) and if R is strictly convex then Π is strictly concave. If, in addition, the optimal actions for the provider are not at the boundaries, then the FOCs are necessary and sufficient to determine the provider's optimal actions. We next present sufficient conditions that guarantee that the maximum is attained at an interior point.

The following conditions imply that $c^*\in(0,c_o)$.

$$\frac{\partial}{\partial c}R(c_o,\mu)>-\lambda(c_o,\mu_o) \text{ and } \frac{\partial}{\partial c}R(0,\mu)<-\Lambda, \text{ for all } \mu\geq\mu_o. \tag{EC47}$$

This follows from (18) because R is convex in c and $\bar\lambda_i\in[\lambda(c_o,\mu_o),\Lambda]$ for all i.

To list conditions that are sufficient for $\mu^* \in (\mu_o, \infty)$ we first introduce some terminology. By (3) and (16) $\frac{t}{(\bar{\mu}_i - \bar{\lambda}_i)} \leq v(\lambda_{min})$ for all $i = 1, \ldots, N$. Hence there exists $K > 0$ such that

$$\frac{\bar{\lambda}_i t}{(\bar{\mu}_i - \bar{\lambda}_i)^2} < K.$$

In addition, this implies that it is never optimal to set $\mu \geq M$ in any equilibrium for a provider by (18) for some $M > 0$. The following conditions are sufficient for μ^* to be an interior point. Assume that there exits μ_K such that $\frac{\partial}{\partial \mu} R(c, \mu) > K$ for $\mu \geq \mu_K$ and for any $c \in [0, c_o]$ and that

$$\frac{\partial}{\partial \mu} R(c, \mu_o) < \frac{t\lambda(c_o, \mu_o)}{(M - \lambda(c_o, \mu_o))^2}, \text{ for all } c \in [0, c_o]. \tag{EC48}$$

The fact that $\mu^* \in (\mu_o, \infty)$ follows from these conditions because; i) R is convex in μ, ii) $\frac{t\lambda}{(\mu - \lambda(c,\mu))^2}$ is minimized at $\mu = \mu_o$ and $c = c_o$ and (iii) M is an upper bound for μ.

EC.3.4. Assumptions for Proposition 4: Regulator's Second-best Solution

We present sufficient conditions for the FOCs to be necessary and sufficient to obtain the regulators' optimal (second-best) actions.

Let $\tilde{c}(\mu)$ be the solution of

$$\frac{\partial}{\partial c} R(\tilde{c}(\mu), \mu) = -\lambda(\mu) \text{ for } \mu \geq \mu_o. \tag{EC49}$$

We show that $\tilde{c}(\mu)$ is well-defined in the proof of Proposition EC7 below. We first define Conditions 6–9 that are used in the next proposition.

Condition 6: $\lambda(\mu) < -\frac{\partial R(0,\mu)}{\partial c}$ for $\mu \geq \mu_o$,
Condition 7: $\lambda(\mu) > -\frac{\partial R(c_o,\mu)}{\partial c}$ for $\mu \geq \mu_o$,
Condition 8: $\left(\frac{t\lambda(\mu)}{(\mu-\lambda(\mu))^2}\left(1 - \frac{\partial \lambda(\mu)}{\partial \mu}\right) - \tilde{c}(\mu)\frac{\partial \lambda(\mu)}{\partial \mu} - \frac{\partial}{\partial \mu} R(\tilde{c}(\mu), \mu)\right)$ is decreasing for $\mu \geq \mu_o$,
Condition 9: $\left(\frac{t\lambda(\mu_o)}{(\mu_o-\lambda(\mu_o))^2}\left(1 - \frac{\partial \lambda(\mu_o)}{\partial \mu}\right) - \tilde{c}(\mu_o)\frac{\partial \lambda(\mu_o)}{\partial \mu}\right) - \frac{\partial}{\partial \mu} R(\tilde{c}(\mu_o), \mu_o) > 0.$

PROPOSITION EC7. *Conditions 6–9 are sufficient for the FOCs of the regulator's problem to be necessary and sufficient to obtain the unique second-best solution.*

Conditions 6, 7, and 9 are the boundary conditions that guarantee the second-best outcomes are interior – similar to Conditions 2, 4, and 5 guaranteeing interior first-best outcomes. In addition, Condition 8 ensures that if $p = 0$ then the welfare function is concave in capacity μ when the provider chooses the optimal marginal cost level $\tilde{c}(\mu)$ for all $\mu \geq \mu_o$. Sufficient conditions presented for the first-best (Conditions 1–5) and second-best (Conditions 6–9) solutions are quite different, despite the fact that the second-best setting can be considered as a special case of the first best. In fact, only Condition 6 implies Condition 4 because $\lambda^*(0, \mu) \leq \lambda(\mu)$ by (EC33).

Proof of Proposition EC7: The proof is similar to that of Proposition EC6 in that we first show that for fixed μ there is a unique optimal cost per patient $\tilde{c}(\mu)$ and then we show that there is a unique optimal μ.

If customers are not charged for service, that is, $p=0$, given any $c \in (0, c_o]$ and $\mu \geq \mu_o$, the welfare function is $S(c,\mu)$, where S is as given in (7) for $p=0$ under the $M/M/1$ assumption. We start by showing that FOCs are sufficient to obtain the second-best marginal cost (denoted by $\tilde{c}(\mu)$) for fixed $\mu \geq \mu_o$. We prove this by showing that $S(\cdot,\mu)$ is concave for fixed μ and $\tilde{c}(\mu)$ cannot be at the boundaries. By (7) (with $p=0$), we have

$$\frac{\partial S(c,\mu)}{\partial c} = -\lambda(\mu) - \frac{\partial R(c,\mu)}{\partial c}, \quad \frac{\partial^2 S(c,\mu)}{\partial c^2} = -\frac{\partial^2 R(c,\mu)}{\partial c^2}. \tag{EC50}$$

Because R is convex $S(\cdot,\mu)$ is concave for fixed μ. Since, in addition, $\frac{\partial S(0,\mu)}{\partial c} > 0$ and $\frac{\partial S(c_o,\mu)}{\partial c} < 0$ by Conditions 6 and 7, given any $\mu \geq \mu_o$ there exists a unique $\tilde{c}(\mu) \in (0, c_o)$ that is obtained by the FOCs and it satisfies (EC49).

We next show that the FOCs are sufficient to obtain the maximum of $S(\tilde{c}(\mu),\mu)$ by showing that it is concave and the optimal is attained at an interior point. By (7) and (EC8), we have

$$\begin{aligned}\frac{d}{d\mu}S(\tilde{c}(\mu),\mu) &= \left(v(\lambda(\mu)) - \frac{t}{\mu-\lambda(\mu)}\right)\frac{\partial\lambda(\mu)}{\partial\mu} - \left(\lambda(\mu) + \frac{\partial R(\tilde{c}(\mu),\mu)}{\partial c}\right)\frac{\partial\tilde{c}(\mu)}{\partial\mu} \\ &\quad + \frac{t\lambda(\mu)}{(\mu-\lambda(\mu))^2}\left(1 - \frac{\partial\lambda(\mu)}{\partial\mu}\right) - \tilde{c}(\mu)\frac{\partial\lambda(\mu)}{\partial\mu} - \frac{\partial}{\partial\mu}R(\tilde{c}(\mu),\mu) \\ &= \frac{t\lambda(\mu)}{(\mu-\lambda(\mu))^2}\left(1 - \frac{\partial\lambda(\mu)}{\partial\mu}\right) - \tilde{c}(\mu)\frac{\partial\lambda(\mu)}{\partial\mu} - \frac{\partial}{\partial\mu}R(\tilde{c}(\mu),\mu),\end{aligned} \tag{EC51}$$

where (EC51) follows from (3) and (EC49). By (EC51) and Condition 8, $\frac{d}{d\mu}S(\tilde{c}(\mu),\mu)$ is decreasing in μ and hence $S(\tilde{c}(\mu),\mu)$ is concave. By (EC51) and Condition 9, $\frac{d}{d\mu}S(\tilde{c}(\mu_o),\mu_o) > 0$. Also, because $\lim_{\mu\to\infty}\frac{t\lambda(\mu)}{(\mu-\lambda(\mu))^2} = 0$, $\frac{d\lambda(\mu)}{d\mu} \geq 0$ by (3) and $\tilde{c}(\mu) > 0$ as discussed above, we have $\lim_{\mu\to\infty}\frac{d}{d\mu}S(\tilde{c}(\mu),\mu) \leq \lim_{\mu\to\infty} -\frac{\partial}{\partial\mu}R(\tilde{c}(\mu),\mu) < 0$. Hence $\mu_o^* \in (\mu_o, \infty)$. Clearly $(\tilde{c}(\mu_o^*), \mu_o^*)$ maximizes S and is the unique solution that satisfies the FOCs under Conditions 6–9. □

EC.3.5. Assumptions for Theorem 2

We provide sufficient conditions that guarantee the FOCs of the providers' profit maximization problem to be necessary and sufficient under the payment scheme of Theorem 2, which yields the profit function in (24).

We claim that if

i) $c\lambda(\mu) + R(c,\mu)$ is jointly convex in c and μ,

ii) v is convex,

iii) $\frac{1}{v}$ is convex,

then $\Pi(c,\mu)$ in (24) is concave. Clearly, if $c\lambda(\mu)+R(c,\mu)$ is convex, it suffices to show that $W(\lambda(\mu),\mu)$ is convex to prove the the concavity of $\Pi(c,\mu)$. By (2)

$$\frac{d^2}{d\mu^2}W(\lambda(\mu),\mu)=(\lambda'(\mu))^2v''(\lambda(\mu))+\lambda''(\mu)v'(\lambda(\mu)). \tag{EC52}$$

Because $v'(\lambda(\mu))<0$ by definition, it is enough to show that $\lambda''(\mu)<0$ to prove the concavity of W. We note that $\frac{1}{v}$ implies $\lambda''(\mu)<0$ because $\lambda(\mu)=\mu-\frac{t}{v(\lambda(\mu))}$ by (2). Hence, if the optimal is attained at an interior point (i)–(iii) guarantee that the FOCs are necessary and sufficient to determine the provider's profit-maximizing actions. We next provide sufficient conditions that imply that this is the case.

Similar to (EC47) assume that

$$\frac{\partial}{\partial c}R(c_o,\mu)>-\lambda(\mu_o) \text{ and } \frac{\partial}{\partial c}R(0,\mu)<-\Lambda, \text{ for all } \mu\geq\mu_o. \tag{EC53}$$

Conditions (EC53) imply that $c_o^*\in(0,c_o)$ by (24) because R is convex in c and $\lambda_i\in[\lambda(\mu_o),\Lambda]$ for all $i=1,\dots,N$.

The following conditions imply that $\mu_o^*\in(\mu_o,\infty)$ by (24) because R is convex in μ. Assume that

$$\frac{\partial}{\partial\mu}R(c,\mu_o)<\left(-c-\lambda(\mu_o)v'(\lambda(\mu_o))\right)\lambda'(\mu_o) \tag{EC54}$$

and

$$\lim_{\tilde{\mu}\to\infty}\frac{\partial}{\partial\mu}R(c,\tilde{\mu})>\lim_{\tilde{\mu}\to\infty}-\left(\Lambda v'(\lambda(\tilde{\mu}))\lambda'(\tilde{\mu})\right) \tag{EC55}$$

for all $c\in[0,c_o]$, where v is defined as in (EC9).

EC.4. Alternative Payment Scheme for First-best with a Unique Equilibrium

In this section we first present an alternative scheme that achieves first-best outcomes in the unique symmetric equilibrium. Then we show that a modified version of this scheme leads to a unique equilibrium, that is, the unique symmetric equilibrium that achieves first-best outcomes is the only equilibrium.

For $\lambda,\mu\geq 0$ let $\Psi_i:\mathbb{R}_+^2\to\mathbb{R}$ be defined as

$$\Psi_i(\lambda,\mu)=\begin{cases}W(\lambda,\mu), \text{ if } \mu\geq\lambda,\\ \bar{W}_i+\frac{\bar{c}_i\lambda+\bar{R}_i}{t\lambda}, \text{ if } \mu<\lambda,\end{cases} \tag{EC56}$$

for given $\bar{W}_i,\bar{c}_i$ and $\bar{R}_i$ and also let the transfer payment, T_i, to provider i be

$$T_i=\bar{R}_i+(\bar{c}_i-c_i)\bar{\lambda_i}-t\lambda_i^2\frac{\partial}{\partial\lambda}W(\lambda_i,\mu_i)-t\bar{\lambda_i}(\Psi_i(\bar{\lambda_i},\mu_i)-\overline{W_i}). \tag{EC57}$$

If, in addition, the toll is set as in (16) the ith provider's objective becomes

$$\Pi(c_i,\mu_i|p_i,T_i) = (\bar{c}_i - c_i)\bar{\lambda}_i - t\bar{\lambda}_i(\Psi_i(\bar{\lambda}_i,\mu_i) - \overline{W_i}) - R(c_i,\mu_i) + \bar{R}_i. \tag{EC58}$$

Before we establish the equilibrium under this payment system, we note that, for a given $\bar{\lambda}_i$ provider i will never choose $\mu_i \le \bar{\lambda}_i$ because by (EC56) its profit will be negative, and it can always set $c_i = \bar{c}_i$ and $\mu_i = \bar{\mu}_i$ to obtain zero profits. Therefore we assume, without loss of generality, that $\mu_i \ge \bar{\lambda}_i$. Also if $\mu_i \ge \bar{\lambda}_i$

$$\Pi(c_i,\mu_i|p_i,T_i) = (\bar{c}_i - c_i)\bar{\lambda}_i - t\bar{\lambda}_i(W(\bar{\lambda}_i,\mu_i) - \bar{W}_i) - R(c_i,\mu_i) + \bar{R}_i \tag{EC59}$$

by (EC56) and (24). For the rest of this section we assume that FOCs are necessary and sufficient for determining the optimal actions of the provider and the waiting time function satisfies the following assumption.

Assumption 1. *We assume that* $\frac{\partial^2}{\partial\lambda^2}W(\lambda,\mu) > 0$, $\frac{\partial^2}{\partial\mu^2}W(\lambda,\mu) > 0$, *and* $\frac{\partial^2 W(\lambda,\mu)}{\partial\mu\partial\lambda} \le 0$.

Proposition EC8. *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* (EC57)*, 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.*

Proof of Proposition EC8: Provider i's optimal actions are obtained by

$$\frac{\partial}{\partial c_i}\Pi(c_i,\mu_i|p_i,T_i) = -\bar{\lambda}_i - \frac{\partial}{\partial c_i}R(c_i,\mu_i) = 0, \tag{EC60}$$

$$\frac{\partial}{\partial \mu_i}\Pi(c_i,\mu_i|p_i,T_i) = -\frac{\partial}{\partial \mu_i}R(c_i,\mu_i) - t\bar{\lambda}_i\frac{\partial}{\partial \mu_i}W(\bar{\lambda}_i,\mu_i) = 0. \tag{EC61}$$

Because first-best marginal cost, c^*, and capacity, μ^*, satisfy (10)–(12), the case where $c_i = c^*$ and $\mu_i = \mu^*$ for all $i = 1,\ldots,N$ is clearly an equilibrium. Also, similar to the proof of Theorem 1, there cannot exist a different symmetric solution because c^* and μ^* are the unique solution to (10)–(12). Hence, there exists a unique symmetric equilibrium that yields first-best outcomes. □

Next we show that if $N = 2$ there cannot be an asymmetric equilibrium.

Proposition EC9. *Assume that $N = 2$ and $\frac{\partial^2 R(c,\mu)}{\partial c\partial\mu} \ge 0$. Under the payment scheme in Proposition EC8 there exists a unique equilibrium.*

We need the following result to prove Proposition EC9.

Proposition EC10. *Consider the payment scheme in Proposition EC8 and assume that $\frac{\partial^2 R(c,\mu)}{\partial c\partial\mu} \ge 0$ and $N = 2$. Let $(\tilde{c}_i,\tilde{\mu}_i)$ for $i = 1,2$ denote an equilibrium and $\tilde{\lambda}_i = \lambda(p_i,\tilde{\mu}_i)$ be given by* (2) *where p_i is defined as in* (16)*. If $\tilde{\lambda}_1 > \tilde{\lambda}_2 > 0$ then $\tilde{\mu}_1 > \tilde{\mu}_2$ and $\tilde{c}_1 < \tilde{c}_2$.*

Proof of Proposition EC9: Similar to the proof of Proposition EC8, we only consider $\mu_i \geq \bar{\lambda}_i$. By (4) if provider j choses action $(\tilde{c}_j, \tilde{\mu}_j)$ provider i's profit function is

$$\Pi(c_i, \mu_i | p_i, T_i) = (\tilde{c}_j - c_i)\tilde{\lambda}_j - R(c_i, \mu_i) + R(\tilde{c}_j, \tilde{\mu}_j) - t\tilde{\lambda}_j(W(\tilde{\lambda}_j, \mu_i) - W(\tilde{\lambda}_j, \tilde{\mu}_j)). \tag{EC62}$$

We next show that an asymmetric equilibrium does not exist when $N = 2$. Let $(\tilde{\mu}_i, \tilde{c}_i)$ denote the action chosen by provider i and $\tilde{\lambda}_i = \lambda(\tilde{p}_i, \tilde{\mu}_i)$ be given by (2) where p_i is defined as in (16)

Assume that $\tilde{\lambda}_1 = \tilde{\lambda}_2$. By (EC60) and (EC62)

$$\frac{\partial R(\tilde{c}_1, \tilde{\mu}_1)}{\partial c} = -\tilde{\lambda}_2 = -\tilde{\lambda}_1 = \frac{\partial R(\tilde{c}_2, \tilde{\mu}_2)}{\partial c}. \tag{EC63}$$

By convexity of R if $\tilde{\mu}_1 = \tilde{\mu}_2$, then $\tilde{c}_1 = \tilde{c}_2$. Hence, a potential asymmetric equilibrium with $\tilde{\lambda}_1 = \tilde{\lambda}_2$ must have $\tilde{\mu}_1 \neq \tilde{\mu}_2$. Without loss of generality we assume $\tilde{\mu}_1 < \tilde{\mu}_2$. By (2), $\tilde{\lambda}_1 = \tilde{\lambda}_2$ implies $p_1 < p_2$ because $W(\tilde{\lambda}_1, \tilde{\mu}_1) > W(\tilde{\lambda}_2, \tilde{\mu}_2)$ when $\tilde{\mu}_1 < \tilde{\mu}_2$ and $\tilde{\lambda}_1 = \tilde{\lambda}_2$. By (16), $p_1 < p_2$ implies $\tilde{c}_1 < \tilde{c}_2$ because

$$\frac{\partial}{\partial \lambda} W(\tilde{\lambda}_1, \tilde{\mu}_1) \geq \frac{\partial}{\partial \lambda} W(\tilde{\lambda}_2, \tilde{\mu}_2)$$

by Assumption 1. Because $\tilde{c}_1 < \tilde{c}_2$ and $\tilde{\mu}_1 < \tilde{\mu}_2$,

$$\frac{\partial R(\tilde{c}_1, \tilde{\mu}_1)}{\partial c} < \frac{\partial R(\tilde{c}_2, \tilde{\mu}_2)}{\partial c}$$

by convexity of R and $\frac{\partial^2 R(c,\mu)}{\partial c \partial \mu} \geq 0$. Because this contradicts (EC63), an asymmetric equilibrium with $\lambda_1 = \lambda_2$ does not exist.

Next we consider the case $\tilde{\lambda}_1 \neq \tilde{\lambda}_2$ and assume without loss of generality that $\tilde{\lambda}_1 > \tilde{\lambda}_2$. Then, by Proposition EC10 we have $\tilde{\mu}_1 > \tilde{\mu}_2$. By Assumption 1, $\tilde{\lambda}_1 > \tilde{\lambda}_2$ and $\tilde{\mu}_1 > \tilde{\mu}_2$ imply

$$W(\tilde{\lambda}_1, \tilde{\mu}_2) - W(\tilde{\lambda}_1, \tilde{\mu}_1) > W(\tilde{\lambda}_2, \tilde{\mu}_2) - W(\tilde{\lambda}_2, \tilde{\mu}_1). \tag{EC64}$$

By (EC62), if $\Pi(\tilde{c}_1, \tilde{\mu}_1 | p_1, T_1) \geq 0$ then

$$\lambda_2 \left(\tilde{c}_2 - \tilde{c}_1 + t(W(\tilde{\lambda}_2, \tilde{\mu}_2) - W(\tilde{\lambda}_2, \tilde{\mu}_1)) \right) \geq R(\tilde{c}_1, \tilde{\mu}_1) - R(\tilde{c}_2, \tilde{\mu}_2).$$

This with (EC64), implies

$$\tilde{\lambda}_2 \left(\tilde{c}_2 - \tilde{c}_1 + t(W(\tilde{\lambda}_1, \tilde{\mu}_2) - W(\tilde{\lambda}_1, \tilde{\mu}_1)) \right) > R(\tilde{c}_1, \tilde{\mu}_1) - R(\tilde{c}_2, \tilde{\mu}_2). \tag{EC65}$$

Then because $\tilde{\lambda}_1 > \tilde{\lambda}_2$, (EC65) and (EC62) imply $\Pi(\tilde{c}_2, \tilde{\mu}_2 | p_2, T_2) < 0$. Hence, an asymmetric equilibrium where two providers experience different equilibrium arrival rates does not exist. □

Proof of Proposition EC10: Let $(\tilde{c}_i, \tilde{\mu}_i)$ for $i = 1, 2$ denote an equilibrium and $\tilde{\lambda}_i = \lambda(p_i, \tilde{\mu}_i)$ be given by (2). Assume that $\tilde{\lambda}_1 > \tilde{\lambda}_2 > 0$. We first show that $\tilde{\mu}_1 > \tilde{\mu}_2$ by contradiction. We show below that

$$\frac{\partial R(\tilde{c}_1, \tilde{\mu}_1)}{\partial c} < \frac{\partial R(\tilde{c}_2, \tilde{\mu}_2)}{\partial c} \tag{EC66}$$

if $\tilde{\mu}_1 \le \tilde{\mu}_2$. By (EC62), provider i's optimal action in equilibrium satisfies

$$\frac{\partial}{\partial c_i}\Pi(\tilde{c}_i, \tilde{\mu}_i | p_i, T_i) = -\tilde{\lambda}_j - \frac{\partial R(\tilde{c}_i, \tilde{\mu}_i)}{\partial c} = 0, \text{for } j \neq i.$$

Because $\tilde{\lambda}_1 > \tilde{\lambda}_2$, we have

$$\frac{\partial R(\tilde{c}_1, \tilde{\mu}_1)}{\partial c} = -\tilde{\lambda}_2 > -\tilde{\lambda}_1 = \frac{\partial R(\tilde{c}_2, \tilde{\mu}_2)}{\partial c},$$

which contradicts (EC66). Hence, if $\tilde{\lambda}_1 > \tilde{\lambda}_2$, $\tilde{\mu}_1 \le \tilde{\mu}_2$ cannot hold.

Next we prove (EC66) if $\tilde{\mu}_1 \le \tilde{\mu}_2$. If $\tilde{\lambda}_1 > \tilde{\lambda}_2$ and $\tilde{\mu}_1 \le \tilde{\mu}_2$ then by (2)

$$(p_1 + tW(\tilde{\lambda}_1, \tilde{\mu}_1)) < (p_2 + tW(\tilde{\lambda}_2, \tilde{\mu}_2)), \tag{EC67}$$

where p_1 and p_2 are given in (16). Because $\tilde{\lambda}_1 > \tilde{\lambda}_2$ and $\tilde{\mu}_1 \le \tilde{\mu}_2$, we have $W(\tilde{\lambda}_1, \tilde{\mu}_1) > W(\tilde{\lambda}_2, \tilde{\mu}_2)$. This implies $p_1 < p_2$ by (EC67). By (16), $p_1 < p_2$ implies

$$\tilde{c}_1 + t\tilde{\lambda}_1 \frac{\partial}{\partial \lambda} W(\tilde{\lambda}_1, \tilde{\mu}_1) < \tilde{c}_2 + t\tilde{\lambda}_2 \frac{\partial}{\partial \lambda} W(\tilde{\lambda}_2, \tilde{\mu}_2). \tag{EC68}$$

By Assumption 1 and that $\tilde{\lambda}_1 > \tilde{\lambda}_2$ and $\tilde{\mu}_1 \le \tilde{\mu}_2$, we have $\frac{\partial W(\tilde{\lambda}_1, \tilde{\mu}_1)}{\partial \lambda} > \frac{\partial W(\tilde{\lambda}_2, \tilde{\mu}_2)}{\partial \lambda}$. This and $\tilde{\lambda}_1 > \tilde{\lambda}_2$ imply that if $p_1 < p_2$ then $\tilde{c}_1 < \tilde{c}_2$ by (EC68). Because $\tilde{\mu}_1 \le \tilde{\mu}_2$ and $\tilde{c}_1 < \tilde{c}_2$ (EC66) holds by convexity of R and $\frac{\partial^2 R(c, \mu)}{\partial c \partial \mu} \ge 0$.

We now show that $\tilde{c}_1 < \tilde{c}_2$ if $\tilde{\lambda}_1 > \tilde{\lambda}_2$. By (EC62) provider i's action satisfies

$$\frac{\partial R(\tilde{c}_i, \tilde{\mu}_i)}{\partial \mu} = -t\tilde{\lambda}_j \frac{\partial}{\partial \mu} W(\tilde{\lambda}_j, \tilde{\mu}_i) \tag{EC69}$$

in any equilibrium. By Assumption 1, and the fact that $\tilde{\lambda}_1 > \tilde{\lambda}_2$ and $\tilde{\mu}_1 > \tilde{\mu}_2$, we have

$$\frac{\partial}{\partial \mu} W(\tilde{\lambda}_1, \tilde{\mu}_2) < \frac{\partial}{\partial \mu} W(\tilde{\lambda}_2, \tilde{\mu}_1).$$

Therefore

$$\left(-\tilde{\lambda}_1 \frac{\partial}{\partial \mu} W(\tilde{\lambda}_1, \tilde{\mu}_2)\right) > \left(-\tilde{\lambda}_2 \frac{\partial}{\partial \mu} W(\tilde{\lambda}_2, \tilde{\mu}_1)\right).$$

Thus by (EC69) we have

$$\frac{\partial R(\tilde{c}_2, \tilde{\mu}_2)}{\partial \mu} > \frac{\partial R(\tilde{c}_1, \tilde{\mu}_1)}{\partial \mu}. \tag{EC70}$$

Also, by the fact that $\mu_1 > \mu_2$ and R is convex

$$\frac{\partial R(\tilde{c}_2, \tilde{\mu}_1)}{\partial \mu} > \frac{\partial R(\tilde{c}_2, \tilde{\mu}_2)}{\partial \mu}. \tag{EC71}$$

By (EC70) and (EC71),

$$\frac{\partial R(\tilde{c}_2, \tilde{\mu}_1)}{\partial \mu} > \frac{\partial R(\tilde{c}_1, \tilde{\mu}_1)}{\partial \mu}.$$

This implies $\tilde{c}_2 > \tilde{c}_1$ by $\frac{\partial^2 R(c,\mu)}{\partial c \partial \mu} \geq 0$. □

We finally consider the case with $N \geq 3$. Assume that providers are divided into two disjoint sets $\mathcal{A}_1$ and $\mathcal{A}_2$. For simplicity, assume that $\mathcal{A}_1 = \{1, 2, \ldots, n_1\}$ and $\mathcal{A}_2 = \{n_1 + 1, \ldots, N\}$, for $n_1 \geq 1$. For $i \in \mathcal{A}_1$, set

$$\bar{c_i}^{\mathcal{A}_1} = \frac{1}{|\mathcal{A}_2|} \sum_{j \in \mathcal{A}_2} c_j,\ \bar{\lambda_i}^{\mathcal{A}_1} = \frac{1}{|\mathcal{A}_2|} \sum_{j \in \mathcal{A}_2} \lambda_j,\ \bar{R_i}^{\mathcal{A}_1} = \frac{1}{|\mathcal{A}_2|} \sum_{j \in \mathcal{A}_2} R_j,\ \bar{W_i}^{\mathcal{A}_1} = \frac{1}{|\mathcal{A}_2|} \sum_{j \in \mathcal{A}_2} W_j.$$

Also set the transfer payment for hospital i equal to T_i defined as in (EC57) using $\bar{c_i}^{\mathcal{A}_1}$, $\bar{\lambda_i}^{\mathcal{A}_1}$, $\bar{W}_i^{\mathcal{A}_1}$ and $\bar{R}_i^{\mathcal{A}_1}$ as defined above. Effectively, we use providers in set $\mathcal{A}_2$ to set the "yardstick" for providers in set $\mathcal{A}_2$ and we can set similar targets for those provider in set $\mathcal{A}_2$ using providers in set $\mathcal{A}_1$.

When providers are split into two sets $\mathcal{A}_1$ and $\mathcal{A}_2$, the providers in the same set would take the same actions since their profit functions are identical. Hence, the comparison of providers in two sets would reduce to the case of $N = 2$, and the unique symmetric equilibrium which yields the first-best outcomes for all providers would be the unique equilibrium by Proposition EC9.

EC.5. Alternative Payment Scheme for Second-best with a Unique Equilibrium

In this section we prove that, under the proposed reimbursement scheme of §4.4, there cannot be an asymmetric equilibrium for $N = 2$. Then, we alter the proposed reimbursement scheme slightly to obtain a new scheme under which the welfare maximising actions are the unique equilibrium.

Proposition EC11. *Assume that there are only two providers (i.e., $N = 2$) and the regulator pays the transfer payment T_i defined as in* (23) *to provider i for $i = 1, 2$. Then there is a unique Nash equilibrium.*

Proof of Proposition EC11: By Theorem 2 the only symmetric equilibrium is where providers pick μ_o^* and c_o^*. We next prove that, for $N = 2$, this is the unique equilibrium. Assume not and let $a_1 = (c_1, \mu_1)$ and $a_2 = (c_2, \mu_2)$ denote another equilibrium with $a_1 \neq a_2$. Under the proposed reimbursement scheme, the profits of providers 1 and 2 are given by

$$\Pi(c_1, \mu_1) = -c_1 \lambda(\mu_1) + t(W_2 - W_1)\lambda(\mu_2) - R(c_1, \mu_1) + R(c_2, \mu_2) + c_2 \lambda(\mu_2). \tag{EC72}$$

$$\Pi(c_2,\mu_2) = -c_2\lambda(\mu_2) + t(W_1 - W_2)\lambda(\mu_2) - R(c_2,\mu_2) + R(c_1,\mu_1) + c_1\lambda(\mu_1). \tag{EC73}$$

In order for (a_1, a_2) to be an equilibrium, we must have $\Pi(c_1,\lambda_1) \geq 0$ and $\Pi(c_2,\lambda_2) \geq 0$ because if $a_2 = a_1$, then $\Pi(c_2,\mu_2) = 0$. We next show that this is not possible in an asymmetric equilibrium. Assume without loss of generality that $\lambda(\mu_1) > \lambda(\mu_2)$. This implies $W_1 < W_2$ by (2), since $p = 0$. By (EC72)

$$t(W_2 - W_1)\lambda(\mu_2) \geq c_1\lambda(\mu_1) + R(c_1,\lambda(\mu_1)) - R(c_2,\lambda(\mu_2)) - c_2\lambda(\mu_2). \tag{EC74}$$

Hence

$$\begin{aligned}\Pi(c_2,\mu_2) = -c_2\lambda(\mu_2) + {} & t(W_1 - W_2)\lambda(\mu_1) - R(c_2,\lambda(\mu_2)) + R(c_1,\lambda(\mu_1)) + c_1\lambda(\mu_1) \\ & \overset{(a)}{<} -c_2\lambda(\mu_2) + t(W_1 - W_2)\lambda(\mu_2) - R(c_2,\lambda(\mu_2)) + R(c_1,\lambda(\mu_1)) + c_1\lambda(\mu_1) \overset{(b)}{\leq} 0,\end{aligned}$$

where (a) above follows from the fact that $W_1 < W_2$ and $\lambda(\mu_1) > \lambda(\mu_2)$, and (b) follows from (EC74). This proves that (a_1, a_2) cannot be an equilibrium if $\lambda_1 \neq \lambda_2$. □

A regulatory scheme with a unique equilibrium for $N \geq 3$: If $N \geq 3$, we can modify the reimbursement scheme so that the only equilibrium is the second best. Specifically, as in Appendix EC.4 we divide the providers into two disjoint sets $\mathcal{A}_1$ and $\mathcal{A}_2$. For simplicity assume that $\mathcal{A}_1 = \{1, 2, \ldots, n_1\}$ and $\mathcal{A}_2 = \{n_1 + 1, \ldots, N\}$ for $n_1 \geq 1$. For $i \in A_1$ set

$$\bar{W}_i^{\mathcal{A}_1} = \frac{1}{|\mathcal{A}_2|}\sum_{j \in A_2} W_j \quad \text{and } \bar{\lambda}_i^{\mathcal{A}_1} = \frac{1}{|\mathcal{A}_2|}\sum_{j \in A_2} \lambda_j.$$

Again, set the transfer payment for hospital $i \in \mathcal{A}_1$ equal to T_i defined as in (23) using $\bar{W}_i^{\mathcal{A}_1}$ and $\bar{\lambda}_i^{\mathcal{A}_1}$ as defined above. Effectively, we use providers in set $\mathcal{A}_2$ to set the "yardstick" for providers in set $\mathcal{A}_2$ and similarly we can use providers in set $\mathcal{A}_2$ to set the "yardstick" for providers in set $\mathcal{A}_2$. The proof that this mechanism can only have a unique equilibrium where each provider picks socially optimal levels μ_o^* and c_o^* follows from a similar argument to that at the end of Appendix EC.4 using Theorem 2 and Proposition EC11. In addition, we conjecture that, for N large enough, the original mechanism should also have a unique equilibrium. This follows from the fact that when N is large, $\bar{W}_i \sim \bar{W}_j$ for all i, j. The proof outlined above for the modified mechanism might be used when this is the case to prove the uniqueness of the symmetric equilibrium.

References

Afanasyev, M., H. Mendelson. 2010. Service provider competition: Delay cost structure, segmentation, and cost advantage. *Manufacturing & Service Operations Management* **12**(2) 213–235.

Armony, M., S. Israelit, A. Mandelbaum, Y. N. Marmor, Y. Tseytlin, G. B. Yom-Tov. 2015. Patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems* **5**(1) 146–194.

Campbell, D. 2016. NHS trust bosses slam £600m hospital fines over patient targets. *The Guardian* (March 29), `https://www.theguardian.com/society/2016/mar/29/nhs-bosses-slam-600m-hospital-fines-over-patient-targets`.

Chen, H., Y-W Wan. 2005. Capacity competition of make-to-order firms. *Oper. Res. Lett.* **33** 187–194.

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

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* **5**(2) 79–141.

Gilboy, N., T. Tanabe, D. Travers, A. M. Rosenau. 2011. Emergency severity index (ESI): A triage tool for emergency department care version 4. Tech. Rep. AHRQ Publication No. 12-0014, Agency for Healthcare Research and Quality, Rockville, MD.

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

McHugh, M., P. Tanabe, M. McClelland, R. K. Khare. 2012. More patients are triaged using the Emergency Severity Index than any other triage acuity system in the United States. *Academic Emergency Medicine* **19**(1) 106–109.

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

ProPublica. 2015. ER Wait Watcher. `https://projects.propublica.org/emergency/`. Accessed: 2015-08-07.

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

Stidham, S. Jr. 2009. *Optimal Design of Queueing Systems*. CRC Press, Boca Raton, FL.