Online Supplement for "Imaging Room and Beyond: The Underlying Economics behind Physicians' Test-Ordering Behavior in Outpatient Services"

Appendix A: Proofs

Proof of Proposition 1. We first show the physician's objective function $g(\mu,p)=p\lambda(\mu,p)$ is concave in p because $\partial^2 g(\mu,p)/\partial p^2=-2p\beta^2\omega/[Q(\mu)-\pi-\beta(p-\pi)]^3-2\beta\omega/[Q(\mu)-\pi-\beta(p-\pi)]^2<0$. Solving the first-order condition gives the optimal service fee p^*, conditional on the service rate μ: $p^*(\mu)=\left\{\mu Q(\mu)-\mu(1-\beta)\pi-\sqrt{\mu\omega[Q(\mu)-(1-\beta)\pi]}\right\}/(\mu\beta)$, using which the physician's objective function can be rewritten as $g(\mu,p^*(\mu))=\left\{r(\mu)+\omega-2\sqrt{r(\mu)\omega}\right\}/\beta$, where $r(\mu)=\mu[Q_c+\alpha\ (\mu_c-\mu)-\pi(1-\beta)]$

Next, we show $g(\mu,p^*(\mu))$ is unimodal in μ. Note that $r(\mu)\geq\mu[Q(\mu)-\pi(1-\beta)]>\mu[Q(\mu)-(\beta p+\pi(1-\beta))]\geq\mu\omega[\mu-\lambda(\mu,p^*(\mu))]^{-1}>\omega$. Hence the sign of

$$dg(\mu,p^*(\mu))/d\mu=(Q_c-\pi(1-\beta)+\alpha(-2\mu+\mu_c))\cdot\left(\sqrt{r(\mu)\omega}-\omega\right)\cdot\left(\beta\sqrt{r(\mu)\omega}\right)^{-1}$$

is the same as that of $Q_c-\pi(1-\beta)+\alpha(-2\mu+\mu_c)$, which is positive when $\mu=0$, decreases in μ, and turns negative when μ is large enough. $g(\mu,p^*(\mu))$ is therefore unimodal in μ. Equating the first-order derivative of $g(\mu,p^*(\mu))$ in terms of μ to zero gives $\mu^*=[Q_c+\alpha\mu_c-(1-\beta)\pi]/(2\alpha)$, which in turn yields $p^*=[Q_c+\alpha\mu_c-(1-\beta)\pi-2\sqrt{\alpha\omega}]/(2\beta)$, and $\lambda^*=\lambda(\mu^*,p^*)=[Q_c+\alpha\mu_c-(1-\beta)\pi]/(2\alpha)-\sqrt{\omega/\alpha}=[Q_c+\alpha\mu_c-(1-\beta)\pi-2\sqrt{\alpha\omega}]/(2\alpha)$. The expected waiting time can thus be determined given μ^* and $\lambda(\mu^*,p^*)$: $W^*=W(\mu^*,\lambda(\mu^*,p^*))=[\mu^*-\lambda(\mu^*,p^*)]^{-1}=\sqrt{\alpha/\omega}$. *Q.E.D.*

Proof of Proposition 2. We first recognize that $U(\mu,\lambda)$ is concave in μ as $\partial^2U(\mu,\lambda)/\partial\mu^2=-2\lambda\omega(\mu-\lambda)^{-3}<0$ for any pair of (μ,λ) that satisfies $\mu>\lambda$. By the first-order condition (in terms of μ), we obtain the conditional expression of the optimal service rate: $\mu^S(\lambda)=\lambda+\sqrt{\omega/\alpha}$, using which the objective function can be rewritten as $-\alpha\lambda^2+(\alpha\mu_c+Q_c-2\sqrt{\alpha\omega})\lambda$, a concave function of λ. The first-order condition gives $\lambda^S=(Q_c+\alpha\mu_c)/(2\alpha)-\sqrt{\omega/\alpha}$, and hence $\mu^S=(Q_c+\alpha\mu_c)/(2\alpha)$. The expected waiting time is thus $W^S=W(\mu^S,\lambda^S)=(\mu^S-\lambda^S)^{-1}=\sqrt{\alpha/\omega}$. *Q.E.D.*

Proof of Proposition 3. Similar to the proof of Proposition 1. *Q.E.D.*

Proof of Corollary 5. We have two cases to consider, depending on the size of $p_{\max}$ (cf. Proposition 3). Case (i): $p_{\max}>\tilde{p}$. In this case, we have $\mu^*=[Q_c+\alpha\mu_c-(1-\beta)\pi]/(2\alpha)<\mu^S=(Q_c+\alpha\mu_c)/(2\alpha)$. Case (ii): $p_{\max}\leq\tilde{p}$. Let $q_{\max}=\pi+\beta(p_{\max}-\pi)$. We have $q_{\max}\leq[Q_c+\alpha\mu_c-2\sqrt{\omega\alpha}+(1-\beta)\pi]/2$. We then divide Case (ii) into two subcases depending on the size of $q_{\max}$: (a) If $(Q_c+\alpha\mu_c-2\sqrt{\alpha\omega})/2<q_{\max}\leq[Q_c+\alpha\mu_c-2\sqrt{\omega\alpha}+(1-\beta)\pi]/2$, then $\mu^*=(Q_c+\alpha\mu_c-q_{\max}-\sqrt{\alpha\omega})/\alpha<(Q_c+\alpha\mu_c)/(2\alpha)=\mu^S$; (b) If $q_{\max}\leq[Q_c+\alpha\mu_c-2\sqrt{\alpha\omega}]/2$, then $\mu^*=(Q_c+\alpha\mu_c-q_{\max}-\sqrt{\alpha\omega})/\alpha\geq(Q_c+\alpha\mu_c)/(2\alpha)=\mu^S$. *Q.E.D.*

Proofs of Propositions 4-5. Similar to the proof of Proposition 1. *Q.E.D.*

Analysis for Section 4: Characterization of the Optimal Scheduling Rule. We first define the set of admissible scheduling rules:

DEFINITION A1. The set Φ of all admissible scheduling rules consists of the nonanticipative, nonpreemptive, and nonidling policies.

We use the achievable region approach (cf. Bertsimas 1996; Federgruen and Groenevelt 1986; Shanthikumar and Yao 1992) to characterize the set of all combinations of waiting times (W_H^ϕ, W_L^ϕ) that can be realized by admissible scheduling rules. The following lemma describes the standard conservation law they need to obey.

LEMMA A1. *A vector of mean waiting times (W_H, W_L) is achievable under an admissible scheduling rule if and only if it satisfies the linear inequalities*

$$W_i \geq \frac{\lambda_H/\mu_H^2 + \lambda_L/\mu_L^2}{1-\lambda_H/\mu_H - \lambda_L/\mu_L} + \frac{1}{\mu_i} \quad \textit{for } i = H, L, \tag{A1}$$

and the conservation law

$$\frac{\lambda_H}{\mu_H} W_H + \frac{\lambda_L}{\mu_L} W_L = \frac{\lambda_H/\mu_H^2 + \lambda_L/\mu_L^2}{1-\lambda_H/\mu_H - \lambda_L/\mu_L}. \tag{A2}$$

The next theorem provides the optimal scheduling policy $\phi \in \Phi$ that minimizes the total expected waiting cost per unit of time $H := \omega\lambda W = \omega\lambda\left(q_H W_H + q_L W_L\right)$. For a proof, see Gelenbe and Mitrani (2010, Chapter 8, Section 3).

THEOREM 1. *The total waiting cost per unit of time is minimized among all admissible scheduling rules $\phi \in \Phi$ by the (non-preemptive)* ***SEPT*** *(*shortest expected processing time*) scheduling policy. That is, strict priority is given to the patient classes in decreasing order of their diagnostic precision α_i.*

Prior to proving Proposition 6, we establish Lemmas A2–A6. For ease of exposition, we write a patient's expected waiting time as a function of λ and Q:

$$W(Q,\lambda) = \frac{q_H/[\mu_H(Q)]^2 + q_L/[\mu_L(Q)]^2}{1/\lambda - q_H/\mu_H(Q)} \cdot \left\{ q_H + \frac{q_L}{1-\lambda\left[q_H/\mu_H(Q) + q_L/\mu_L(Q)\right]} \right\} + \frac{q_H}{\mu_H(Q)} + \frac{q_L}{\mu_L(Q)},$$

and define

$$f_H = \frac{q_H/[\mu_H(Q)]^2 + q_L/[\mu_L(Q)]^2}{1/\lambda - q_H/\mu_H(Q)} \text{ and } f_L = q_H + \frac{q_L}{1-\lambda\left[q_H/\mu_H(Q) + q_L/\mu_L(Q)\right]}.$$

Clearly, we have $\partial f_H/\partial\lambda > 0, \partial f_H/\partial Q > 0, \partial f_L/\partial\lambda > 0$, and $\partial f_L/\partial Q > 0$.

LEMMA A2. *$W(Q,\lambda)$ increases strictly with the service quality Q.*

Proof of Lemma A2. This can be shown by noting that

$$\begin{aligned}\frac{\partial W(Q,\lambda)}{\partial Q} &= \frac{\partial W(Q,\lambda)}{\partial \mu_H(Q)} \cdot \mu_H'(Q) + \frac{\partial W(Q,\lambda)}{\partial \mu_L(Q)} \cdot \mu_L'(Q) \\ &= \frac{\partial W(Q,\lambda)}{\partial \mu_H(Q)} \cdot \left(-\frac{1}{\alpha_H}\right) + \frac{\partial W(Q,\lambda)}{\partial \mu_L(Q)} \cdot \left(-\frac{1}{\alpha_L}\right) > 0.\end{aligned}$$

Q.E.D.

LEMMA A3. *The effect of each variable (λ and Q) on the expected waiting time $W(Q,\lambda)$ is more salient as the other increases, that is, $\partial^2 W(Q,\lambda)/\partial Q\partial\lambda > 0$.*

Proof of Lemma A3. We first note that

$$\begin{aligned}\frac{\partial^2 f_H}{\partial Q \partial \lambda} &= \frac{\partial^2 f_H}{\partial \lambda \partial \mu_H(Q)} \cdot \mu'_H(Q) + \frac{\partial^2 f_H}{\partial \lambda \partial \mu_L(Q)} \cdot \mu'_L(Q) \\ &= \left\{ -\frac{2q_H(\mu_H(Q)\lambda q_L + [\mu_L(Q)]^2)}{[\mu_L(Q)]^2(\mu_H(Q) - \lambda q_H)^2} \right\} \cdot \left(-\frac{1}{\alpha_H} \right) + \left\{ -\frac{2q_L[\mu_H(Q)]^2}{[\mu_L(Q)]^3(\mu_H(Q) - \lambda q_H)^2} \right\} \cdot \left(-\frac{1}{\alpha_L} \right) > 0\end{aligned}$$

and

$$\begin{aligned}\frac{\partial^2 f_L}{\partial Q \partial \lambda} =& \frac{\partial^2 f_L}{\partial \lambda \partial \mu_H(Q)} \cdot \mu'_H(Q) + \frac{\partial^2 f_L}{\partial \lambda \partial \mu_L(Q)} \cdot \mu'_L(Q) \\ =& \left\{ -\frac{q_H q_L \{[\mu_L(Q)]^2 [1 - \lambda(q_H/\mu_H(Q) + q_L/\mu_L(Q)]\}}{\{1 - \lambda[q_H/\mu_H(Q) + q_L/\mu_L(Q)]\}^3} \right\} \cdot \left(-\frac{1}{\alpha_H} \right) \\ &+ \left\{ -\frac{q_H q_L [\mu_H(Q)]^2 \{1 - \lambda[q_H/\mu_H(Q) + q_L/\mu_L(Q)]\}}{[1 - \lambda(q_H/\mu_H(Q) + q_L/\mu_L(Q))]^3} \right\} \cdot \left(-\frac{1}{\alpha_L} \right) > 0.\end{aligned}$$

Therefore,

$$\frac{\partial^2 W(Q,\lambda)}{\partial Q \partial \lambda} = \frac{\partial^2 (f_H \cdot f_L)}{\partial Q \partial \lambda} = f_H \frac{\partial^2 f_L}{\partial Q \partial \lambda} + f_L \frac{\partial^2 f_H}{\partial Q \partial \lambda} + \frac{\partial f_H}{\partial \lambda} \cdot \frac{\partial f_L}{\partial Q} + \frac{\partial f_H}{\partial Q} \cdot \frac{\partial f_L}{\partial \lambda} > 0.$$

Q.E.D.

Lemma A4. *$W(Q,\lambda)$ is convex in λ.*

Proof of Lemma A4. Note that

$$\begin{aligned}\frac{\partial^2 f_H}{\partial \lambda^2} &= \frac{2q_H\{q_H/[\mu_H(Q)]^2 + q_L/[\mu_L(Q)]^2\}}{\mu_H(Q)[1 - q_H\lambda/\mu_H(Q)]^3} > 0, \text{ and} \\ \frac{\partial^2 f_L}{\partial \lambda^2} &= \frac{2q_L[q_L\mu_H(Q) + q_H\mu_L(Q)]^2}{[1 - q_H\lambda/\mu_H(Q) - q_L\lambda/\mu_L(Q)]^3} > 0,\end{aligned}$$

which gives

$$\frac{\partial^2 W(Q,\lambda)}{\partial \lambda^2} = \frac{\partial^2 (f_H \cdot f_L)}{\partial \lambda^2} = f_H \frac{\partial^2 f_L}{\partial \lambda^2} + f_L \frac{\partial^2 f_H}{\partial \lambda^2} + 2\frac{\partial f_H}{\partial \lambda} \cdot \frac{\partial f_L}{\partial \lambda} > 0.$$

Q.E.D.

Lemma A5. *$\lambda W(Q,\lambda)$ is convex in λ.*

Proof of Lemma A5. We have

$$\frac{\partial^2 [\lambda W(Q,\lambda)]}{\partial \lambda^2} = \lambda \frac{\partial^2 W(Q,\lambda)}{\partial \lambda^2} + 2\frac{\partial W(Q,\lambda)}{\partial \lambda} > 0$$

from Lemma A4. *Q.E.D.*

Lemma A6. *$W(Q,\lambda)$ is convex in Q.*

Proof of Lemma A6. We first establish the following results:

$$\begin{aligned}\frac{\partial f_H}{\partial \mu_H(Q)} &= -\frac{q_H\lambda\{q_L\lambda[\mu_H(Q)]^2 + [2\mu_H(Q) - q_H\lambda][\mu_L(Q)]^2\}}{[\mu_H(Q)\mu_L(Q)]^2[\mu_H(Q) - q_H\lambda]^2} < 0, \\ \frac{\partial f_H}{\partial \mu_L(Q)} &= -\frac{2q_L\lambda}{[1 - q_H\lambda/\mu_H(Q)][\mu_L(Q)]^3} < 0, \\ \frac{\partial f_L}{\partial \mu_H(Q)} &= -\frac{q_H q_L \lambda}{[\mu_H(Q)]^2[1 - q_H\lambda/\mu_H(Q) - q_L\lambda/\mu_L(Q)]^2} < 0,\end{aligned}$$

$$\frac{\partial f_L}{\partial \mu_L(Q)} = -\frac{q_L^2\lambda}{[\mu_L(Q)]^2[1-q_H\lambda/\mu_H(Q)-q_L\lambda/\mu_L(Q)]^2} < 0,$$
$$\frac{\partial^2 f_H}{\partial[\mu_H(Q)]^2} = \frac{2q_H\lambda\{q_L\lambda[\mu_H(Q)]^3+q_H^2\lambda^2[\mu_L(Q)]^2+3\mu_H(Q)[\mu_H(Q)-q_H\lambda]\}}{\{\mu_H(Q)[\mu_H(Q)-q_H\lambda]\}^3[\mu_L(Q)]^2} > 0,$$
$$\frac{\partial^2 f_H}{\partial[\mu_L(Q)]^2} = \frac{6q_L\lambda}{[1-q_H\lambda/\mu_H(Q)][\mu_L(Q)]^4} > 0,$$
$$\frac{\partial^2 f_H}{\partial\mu_H(Q)\partial\mu_L(Q)} = \frac{2q_Hq_L\lambda^2}{[\mu_H(Q)-q_H\lambda]^2[\mu_L(Q)]^3} > 0,$$
$$\frac{\partial^2 f_L}{\partial[\mu_H(Q)]^2} = \frac{2q_Hq_L\lambda[\mu_L(Q)-q_L\lambda]}{[\mu_H(Q)]^2[1-q_H\lambda/\mu_H(Q)-q_L\lambda/\mu_L(Q)]^3} > 0,$$
$$\frac{\partial^2 f_L}{\partial[\mu_L(Q)]^2} = \frac{2q_L^2\lambda[\mu_H(Q)-q_H\lambda]}{[\mu_L(Q)]^2[1-q_H\lambda/\mu_H(Q)-q_L\lambda/\mu_L(Q)]^3} > 0, \text{ and}$$
$$\frac{\partial^2 f_L}{\partial\mu_H(Q)\partial\mu_L(Q)} = \frac{2q_Hq_L^2\lambda^2}{[\mu_H(Q)\mu_L(Q)]^2[1-q_H\lambda/\mu_H(Q)-q_L\lambda/\mu_L(Q)]^3} > 0.$$

These results yield

$$\frac{\partial^2(f_H\cdot f_L)}{\partial[\mu_H(Q)]^2} = f_H\frac{\partial^2 f_L}{\partial[\mu_H(Q)]^2} + f_L\frac{\partial^2 f_H}{\partial[\mu_H(Q)]^2} + 2\frac{\partial f_H}{\partial\mu_H(Q)}\cdot\frac{\partial f_L}{\partial\mu_H(Q)} > 0,$$
$$\frac{\partial^2(f_H\cdot f_L)}{\partial[\mu_L(Q)]^2} = f_H\frac{\partial^2 f_L}{\partial[\mu_L(Q)]^2} + f_L\frac{\partial^2 f_H}{\partial[\mu_L(Q)]^2} + 2\frac{\partial f_H}{\partial\mu_L(Q)}\cdot\frac{\partial f_L}{\partial\mu_L(Q)} > 0, \text{ and}$$
$$\frac{\partial^2(f_H\cdot f_L)}{\partial\mu_H(Q)\partial\mu_L(Q)} = f_H\frac{\partial^2 f_L}{\partial\mu_H(Q)\partial\mu_L(Q)} + f_L\frac{\partial^2 f_H}{\partial\mu_H(Q)\partial\mu_L(Q)} + \frac{\partial f_H}{\partial\mu_H(Q)}\cdot\frac{\partial f_L}{\partial\mu_L(Q)} + \frac{\partial f_H}{\partial\mu_L(Q)}\cdot\frac{\partial f_L}{\partial\mu_H(Q)} > 0.$$

Therefore, we have

$$\frac{\partial^2 W(Q,\lambda)}{\partial Q^2} = \frac{1}{\alpha_H^2}\cdot\frac{\partial^2(f_H\cdot f_L)}{\partial[\mu_H(Q)]^2} + \frac{2}{\alpha_H\alpha_L}\cdot\frac{\partial^2(f_H\cdot f_L)}{\partial\mu_H(Q)\partial\mu_L(Q)} + \frac{1}{\alpha_L^2}\cdot\frac{\partial^2(f_H\cdot f_L)}{\partial[\mu_L(Q)]^2} + \frac{2q_H}{\alpha_H^2[\mu_H(Q)]^3} + \frac{2q_L}{\alpha_L^2[\mu_L(Q)]^3} > 0.$$

Q.E.D.

Proof of Proposition 6. The objective functions for the market equilibrium and the social optimum are $g(Q,\lambda) = \lambda[Q-\omega W(Q,\lambda)-\pi(1-\beta)]$ and $U(Q,\lambda) = \lambda[Q-\omega W(Q,\lambda)]$, respectively. Denote $k := \pi(1-\beta)$. Note that given λ, the optimum Q in both problems are found using the following sub-optimization problem:

$$\max_Q\ Q-\omega W(Q,\lambda). \tag{A3}$$

Let $\Psi(\lambda)$ denote the optimal objective function value of (A3). Use the notation $f(Q,\lambda)$ to denote its objective function, namely,

$$\Psi(\lambda) := \max_Q\ f(Q,\lambda) = Q-\omega W(Q,\lambda).$$

Because W is convex in Q, $\Psi(\lambda)$ exists and is continuous.

Next, consider the optimization problem for λ, which we can write as follows:

$$\max_\lambda\ g(\lambda,k) = \lambda(\Psi(\lambda)-k).$$

In the objective function $g(\lambda,k)$, we are treating k as a parameter and λ as the decision variable. We want to compare the optimal solutions for different values of k. More precisely, let λ^0 maximize $g(\cdot,k^0)$ and λ^1 maximize $g(\cdot,k^1)$. Here, $k^0=0$ and $k^1=\pi(1-\beta)>0$. Therefore, $\lambda^0=\lambda^S$ and $\lambda^1=\lambda^*$.

First note that

$$\frac{\partial^2 g(\lambda,k)}{\partial\lambda\partial k} = -1 < 0.$$

Because λ^0 maximizes $g(\cdot, k^0)$ and λ^1 maximizes $g(\cdot, k^1)$, we have

$$g\left(\lambda^0, k^0\right) \geq g\left(\lambda^1, k^0\right) \text{ and } g\left(\lambda^1, k^1\right) \geq g\left(\lambda^0, k^1\right).$$

Subtract the first inequality from the second to obtain

$$g\left(\lambda^1, k^1\right) - g\left(\lambda^1, k^0\right) \geq g\left(\lambda^0, k^1\right) - g\left(\lambda^0, k^0\right). \tag{A4}$$

Now define the function $h(\lambda)$ of λ:

$$h(\lambda) := g\left(\lambda, k^1\right) - g\left(\lambda, k^0\right).$$

Then, inequality (A4) becomes

$$h\left(\lambda^1\right) \geq h\left(\lambda^0\right). \tag{A5}$$

We can rewrite (A5) as follows:

$$0 \leq h\left(\lambda^1\right) - h\left(\lambda^0\right) = \int_{\lambda^0}^{\lambda^1} h'(\lambda)\, d\lambda = \int_{\lambda^0}^{\lambda^1} \left[\frac{\partial g(\lambda, k^1)}{\partial \lambda} - \frac{\partial g(\lambda, k^0)}{\partial \lambda}\right] d\lambda. \tag{A6}$$

Next, note the integrand can also be written as an integral over k. That is,

$$\frac{\partial g(\lambda, k^1)}{\partial \lambda} - \frac{\partial g(\lambda, k^0)}{\partial \lambda} = \int_{k^0}^{k^1} \frac{\partial^2 g(\lambda, k)}{\partial \lambda \partial k} dk.$$

Substitute this back into (A6):

$$h\left(\lambda^1\right) - h\left(\lambda^0\right) = \int_{\lambda^0}^{\lambda^1} \int_{k^0}^{k^1} \frac{\partial^2 g(\lambda, k)}{\partial \lambda \partial k} dk d\lambda \geq 0.$$

We can also rewrite the objective function under social welfare optimization as follows:

$$U(Q, \lambda) = \lambda \left[Q - \omega W(Q, \lambda)\right].$$

Because $k^1 > k^0$ and $\frac{\partial^2 g(\lambda,k)}{\partial \lambda \partial k} < 0$, the first integral (with respect to k) is negative. Therefore, for the entire integral to be nonnegative, it must be that $\lambda^1 \leq \lambda^0$. (Recall that $\int_a^b = -\int_b^a$.)

Having shown $\lambda^1 \leq \lambda^0$, we show $Q^1 \geq Q^0$ by zooming on the subproblem

$$\max_Q \; f(Q, \lambda) = Q - \omega W(Q, \lambda).$$

Now, in the objective function $f(Q, \lambda)$, we are treating λ as a parameter and Q as the decision variable. Let Q^0 maximize $f(\cdot, \lambda^0)$ and Q^1 maximize $f(\cdot, \lambda^1)$. Because

$$\frac{\partial^2 f(Q, \lambda)}{\partial Q \partial \lambda} = -\frac{\partial^2 W(Q, \lambda)}{\partial Q \partial \lambda} < 0,$$

and $\lambda^1 \leq \lambda^0$, by repeating similar arguments as above, we show that indeed $Q^1 \geq Q^0$.

Finally, we show $W(Q^1, \lambda^1) \leq W(Q^0, \lambda^0)$ when ω is large enough (in a sense we make precise below). Let $Q(\lambda)$ be the optimal quality for a given λ, that is,

$$Q(\lambda) = \arg\max f(Q, \lambda) = Q - \omega W(Q, \lambda).$$

Consider $\frac{dW(Q(\lambda),\lambda)}{d\lambda}$. We can write

$$\frac{dW(Q(\lambda), \lambda)}{d\lambda} = \frac{\partial W(Q(\lambda), \lambda)}{\partial \lambda} + \frac{\partial W(Q(\lambda), \lambda)}{\partial Q} \frac{dQ(\lambda)}{\partial \lambda}. \tag{A7}$$

First, note that the optimal $Q(\lambda)$ satisfies the necessary first order condition for optimality

$$\frac{1}{\omega} = \frac{\partial W(Q,\lambda)}{\partial Q}. \tag{A8}$$

Second, to calculate $\frac{dQ(\lambda)}{\partial \lambda}$, we differentiate the above first order condition implicitly with respect to λ. This yields

$$0 = \frac{\partial^2 W(Q,\lambda)}{\partial Q^2}\frac{dQ(\lambda)}{\partial \lambda} + \frac{\partial W(Q,\lambda)}{\partial \lambda},$$

hence,

$$\frac{dQ(\lambda)}{\partial \lambda} = \frac{-\frac{\partial W(Q,\lambda)}{\partial \lambda}}{\frac{\partial^2 W(Q,\lambda)}{\partial Q^2}}. \tag{A9}$$

Substituting (A8) and (A9) into (A7),

$$\frac{dW(Q(\lambda),\lambda)}{d\lambda} = \frac{\partial W(Q(\lambda),\lambda)}{\partial \lambda} - \frac{1}{\omega}\frac{\frac{\partial W(Q,\lambda)}{\partial \lambda}}{\frac{\partial^2 W(Q,\lambda)}{\partial Q^2}} = \frac{\partial W(Q(\lambda),\lambda)}{\partial \lambda}\left(1 - \frac{\frac{1}{\omega}}{\frac{\partial^2 W(Q,\lambda)}{\partial Q^2}}\right).$$

Since $\partial W(Q(\lambda),\lambda)/\partial \lambda > 0$, we have $dW(Q(\lambda),\lambda)/d\lambda > 0$ if $\omega > \left[\frac{\partial^2 W(Q,\lambda)}{\partial Q^2}\right]^{-1}$. In other words, $dW(Q(\lambda),\lambda)/d\lambda > 0$ if ω is large enough (or if $W(\cdot,\cdot)$ is convex enough) and vice versa. Since $\lambda^1 \le \lambda^0$, this proves that $W(Q(\lambda^1),\lambda^1) \le W(Q(\lambda^0),\lambda^0)$ when ω is large enough (or W is convex enough). *Q.E.D.*

Proof of Proposition 7. Because $\beta' \ge \beta$ and $(1-\beta')\pi' \ge (1-\beta)\pi$, given any service fee p, a patient of the original type always has an out-of-pocket expense that is no higher than a patient of the new type. In other words, for any service-parameter combination (μ, p), by joining the queue, a patient of the original type always receives a higher net expected surplus than a patient of the new type. As a result, for any combination of (μ, p), two outcomes are possible in equilibrium: (1) some type patients of the original type join the queue but no patients of the new type join the queue, and (2) all the patients of the original type join the queue, and some (but not all) patients of the new type join the queue. An outcome cannot sustain where some type patients of the new type join the queue, but not all patients of the original type join the queue; otherwise, given any service rate μ, a higher service fee p can be imposed such that by joining the queue, a patient of the new type would have a negative expected surplus, whereas a patient of the original type has a nonnegative expected surplus, effectively "replacing" a patient of the new type with a patient of the original type.

In the first case, only patients of the original type would join the queue. Following the same procedure as in the proof of Proposition 1, we can derive the optimal service rate $\mu^* = [Q_c + \alpha\mu_c - (1-\beta)\pi]/(2\alpha) = \hat{\mu}$, and the equilibrium arrival rate is $\lambda^* = \hat{\mu} - \sqrt{\omega/\alpha}$. We need $\lambda^* \le \alpha\Lambda$ to ensure the equilibrium arrival rate does not exceed the potential arrival rate of the original type of patients.

In the second case, both types of patients will join the queue. Because a patient of the new type must receive a nonnegative surplus in equilibrium, it follows that a patient of the original type would receive a positive surplus. Therefore, in equilibrium, every patient of the original type joins the queue with a probability of 1 and earns a positive expected surplus, whereas a patient of the new type joins the queue with a probability of less than 1 and earns an expected surplus of zero. In other words, a patient of the original type, has a

non-binding participation constraint. Let λ' be the equilibrium arrival rate of patients of the new type. The physicians' problem can be formulated as

$$\begin{aligned} \max_{\mu,p} \quad & p\cdot(\lambda'+(1-\gamma)\Lambda) \\ \text{subject to} \quad & Q(\mu)-\pi_b-\beta_b(p-\pi_b)-\frac{\omega}{\mu-\lambda'-(1-\gamma)\Lambda}\geq 0, \end{aligned}$$

solving which gives the optimal service rate $\mu^*=\hat{\mu}'$, yielding an induced arrival rate of $\lambda'=\hat{\mu}'-\sqrt{\omega/\alpha}-(1-\gamma)\Lambda$. As a boundary condition, we need $\lambda'+(1-\gamma)\Lambda<\Lambda$, that is, $\hat{\mu}'-\sqrt{\omega/\alpha}<\Lambda$, to ensure the total induced arrival rate does not exceed the total potential arrival rate. This is satisfied due to the assumption that $\Lambda>\hat{\mu}'-\sqrt{\omega/\alpha}$.

In determining the socially optimal service rate, note that in the eyes of the social planner, no difference exists between the two types of patients, because the insurance coverage is irrelevant in determining patients' queue-joining probabilities. Hence, the socially optimal service rate $\mu^{SO}=\frac{Q_c+\alpha\mu_c}{2\alpha}$, which is the same as in the case in which no heterogeneity exists in insurance coverage. *Q.E.D.*

Analysis for §5: The Case with $N\geq 2$ Types of Insurance Coverage. Each type-i patient has an insurance policy with copayment π_i and coinsurance β_i. The proportion of type-i patients is γ_i for $i=1,\cdots,N$, with $\sum_{i=1}^N\gamma_i=1$. The patient type is indexed such that $\beta_H\leq\beta_L\leq\cdots\leq\beta_N$ and $(1-\beta_H)\pi_H\leq(1-\beta_L)\pi_L\leq\cdots(1-\beta_N)\pi_N$. As such, for any service fee p, a type-j patient's out-of-pocket expense is no higher than a type-k patient's, provided that $j<k$. For simplicity of presentation, we define

$$\hat{\mu}_i=\frac{Q_c+\alpha\mu_c-(1-\beta_i)\pi_i}{2\alpha} \text{ for } i=1,\cdots,N$$

as the optimal service rate when all the patients are of type-i. We assume $\Lambda>\hat{\mu}_N-\sqrt{\omega/\alpha}$, meaning the total potential arrival rate is large enough such that not all patients would be covered in equilibrium.

The proposition below states the optimal service rate.

PROPOSITION A1. *In the case of N types of patients, the optimal service rate is $\mu^*=\hat{\mu}_{i^*}$, where $i^*=1$ if $\mu_H-\sqrt{\omega/\alpha}\leq\Lambda_H$; otherwise, $i^*\in\{2,\cdots,N\}$ is the unique integer that satisfies*

$$\mu_{i^*-1}-\sqrt{\omega/\alpha}>\Lambda\cdot\sum_{k=1}^{i^*-1}\gamma_k, \text{ and } \mu_{i^*}-\sqrt{\omega/\alpha}<\Lambda\cdot\sum_{k=1}^{i^*}\gamma_k.$$

Proposition A1 states that a critical type of insurance coverage ($i^*\in\{1,\cdots,N\}$) exists that would become the focal point in the physicians' choice of service parameters. The critical type i^* corresponds to the equilibrium in which all the type-1, type-2, $\cdots$, type-(i^*-1) patients choose to join the queue, a proportion of type-i^* patients join the queue, and none of type-$i^*,\cdots,N$ patients join the queue.

Proof of Proposition A1. We proceed similar to the proof of Proposition 7. Note that the complete condition for $\mu^*=\hat{\mu}_{i^*}$, $i^*\in\{2,\cdots,N\}$ is

$$\mu_j-\sqrt{\omega/\alpha}>\Lambda\cdot\sum_{k=1}^{j}\gamma_k \text{ for all } j=1,\cdots,i^*-1, \text{ and } \mu_{i^*}-\sqrt{\omega/\alpha}<\Lambda\cdot\sum_{k=1}^{i^*}\gamma_k.$$

The first part of the above condition is equivalent to $\mu_{i^*-1}-\sqrt{\omega/\alpha}>\Lambda\cdot\sum_{k=1}^{i^*-1}\gamma_k$ because μ_{j-1} decreases in j while $\Lambda\cdot\sum_{k=1}^{j-1}\gamma_k$ increases in j. Thus, we have $\mu_j-\sqrt{\omega/\alpha}>\Lambda\cdot\sum_{k=1}^{j}\gamma_k$ for all $j=1,\cdots,i^*-1$ from $\mu_{i^*-1}-\sqrt{\omega/\alpha}>\Lambda\cdot\sum_{k=1}^{i^*-1}\gamma_k$. *Q.E.D.*

Appendix B: Robustness Check

To explore the robustness of our main results, we conduct a thorough investigation into various extensions of our baseline model. These extensions are available upon request.

- **A general, non-monotonic service-quality function.** Our model assumes the service quality $Q(\mu)$ to be an affine function of μ for simplicity of analysis. When we extend $Q(\mu)$ to a general concave function and allow it to be non-monotonic, that is, excessive testing leads to a lower service value, we can show that all of our major insights continue to hold.
- **Service-time distribution.** Throughout the paper, we assume the service time is exponentially distributed. We show that all of our results pertaining to the physician's test-ordering behavior remain valid when the service time follows a general distribution. Nevertheless, the coefficient of variation of the general distribution is reflected in the optimal service fee.
- **Definition of waiting time.** We have thus far used the total time in the system as the definition of the waiting time, because the benefits from the service have been captured by the function $Q(\mu)$. If the waiting time is defined as the time in the queue only, which would correspond to patients incurring waiting costs only until they start the process of completing their tests, we can show our major results remain unchanged.
- **Misdiagnosis cost function.** Our model assumes an affine misdiagnosis cost function. When this assumption is relaxed to a general function that is convex increasing in μ, our insights regarding misdiagnosis remains unchanged.
- **Asymmetric information in physician skill level.** We address the issue of information asymmetry about physicians' skill levels, and show that physicians' signaling efforts can lead to more salient overtesting behavior, especially when technological advancements flatten out differentiation among physicians.
- **Follow-up visits.** We assume in our model that at any time, every patient with medical needs adopts the same randomized queue-joining strategy; consequently, a patient can make multiple visits, but different visits are deemed independent of each other. In practice, however, the physician advises a proportion of patients making initial visits to make follow-up visits. We can show that although the physician exhibits different test-ordering patterns for new and returning patients, our major findings remain directionally valid.
- **Labor cost.** Our model ignores the labor costs involved in the analysis of test of results. When the cost of labor is incorporated, we find it influences the optimal service fee but not the optimal service rate.

References

Bertsimas, D.. 1996. The achievable region method in the optimal control of queueing systems. *Queueing Systems* **21**(3–4) 337–389.

Federgruen A., H. Groenevelt. 1988. Characterization and optimization of achievable performance in general queueing systems. *Oper. Res.* **36**(5) 733–741.

Gelenbe, E., I. Mitrani. 2010. *Analysis and Synthesis of Computer Systems. 2nd Ed.*. London, U.K.: Imperial College Press.

Shanthikumar, J. G., D. D. Yao. 1992. Multiclass queueing systems: Polymatroidal structure and optimal scheduling control. *Oper. Res.* **40**(2) 293–299.