Online Appendix

Time-Varying Physician Productivity and Implications for Emergency Department Modeling and Staffing

Zhankun Sun, Ran Liu, Huiyin Ouyang

A. Further results on data analysis

Figure A1 **The average new patients seen per hour (PPH) for 8-hour shifts in the main ED area with a time resolution of 1 hour. The extra point on the curve for physician workload outside the shift duration is due to physician overtime for one hour.**

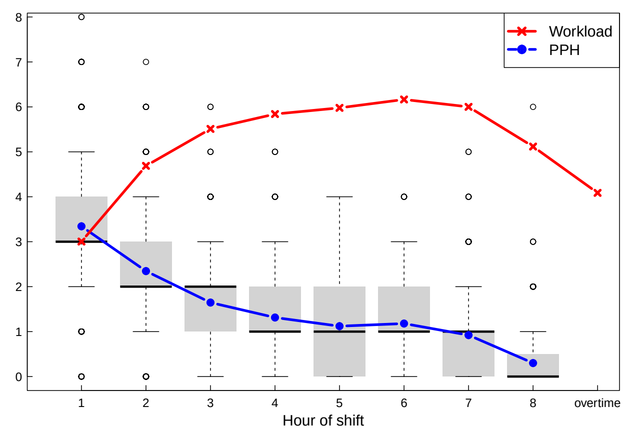


Figure A2 **The average new patients seen per hour (PPH) for 7- and 8-hour shifts with a time resolution of 30 minutes. The extra point for physician workload outside the shift duration is due to physician overtime for one hour.**

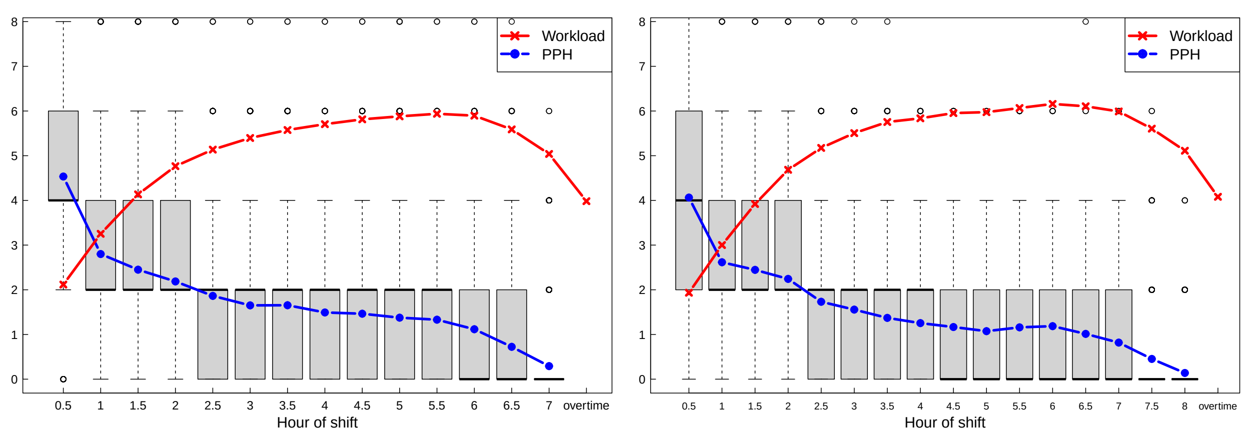
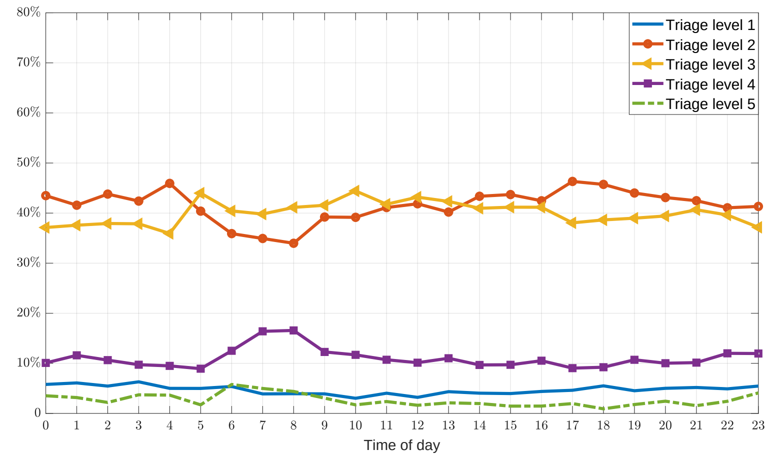


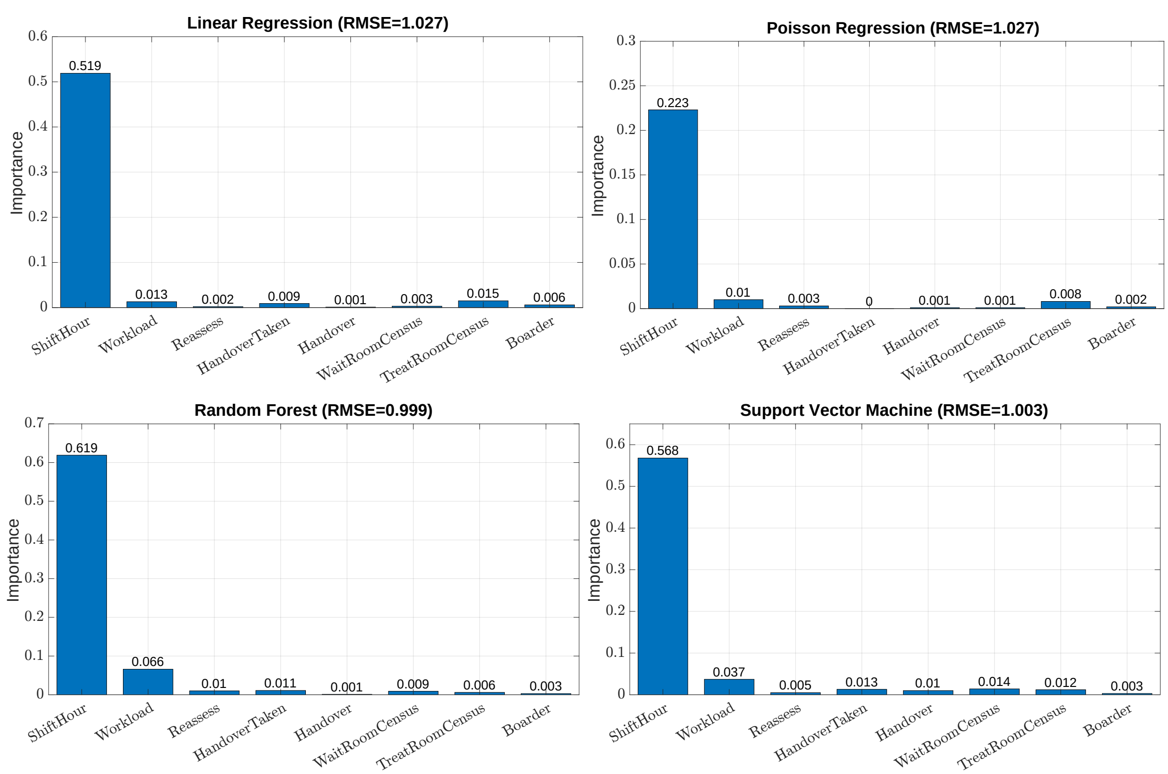
Figure A3 **The percentage of patients from each triage level in the main ED area by the time of day.**



B. Importance of Factors in Predicting PPH

In this section, we explore the importance of these factors from a prediction perspective. By understanding which factors most significantly influence *PPH*, we can better predict physician productivity and potentially identify potential leverage points for operational improvement.

Figure A4 **The feature importance of variables using four models to predict the PPH for 7-hour shifts and their corresponding root mean squared error (RMSE). The physician, shift ID, and weekend are controlled. The feature importance is calculated by permuting the corresponding variable 50 times.**



We employed the permutation importance method (Fisher et al. 2019) to evaluate the significance of each factor in predicting PPH rates for different models. This method assesses feature importance by measuring the increase in the model's prediction error after permuting the feature values while keeping the other features unchanged. The resulting increase in the prediction error indicates the significance of the feature in predicting the target variable.

In our analysis, we use the root mean squared error (RMSE) as the model performance measure. We experiment with several prediction models, including linear regression, Poisson regression, random forest, and support vector machine (SVM). All variables described in Section 3.2 are included in the prediction models. We randomly divide the dataset into training (60%) and testing (40%) datasets. For each model, we first train the model on the training dataset and then predict PPH using the testing dataset to obtain the RMSE. We then randomly permute the feature of interest multiple times on the testing dataset and re-evaluate the RMSE. The importance of the feature is determined by the average increase in the RMSE before and after permuting the feature values.

Figures A4 and A5 present the increase in RMSE when each feature is randomly permuted 50 times while keeping others unchanged. Across all four regression models, it is evident that *ShiftHour* is the most important factor in predicting PPH, while *WaitRoomCensus*, *TreatRoomCensus*, and *Boarder* exhibit only marginal effects. *Workload* appears to be more important than the census measures but less important compared to *ShiftHour*.

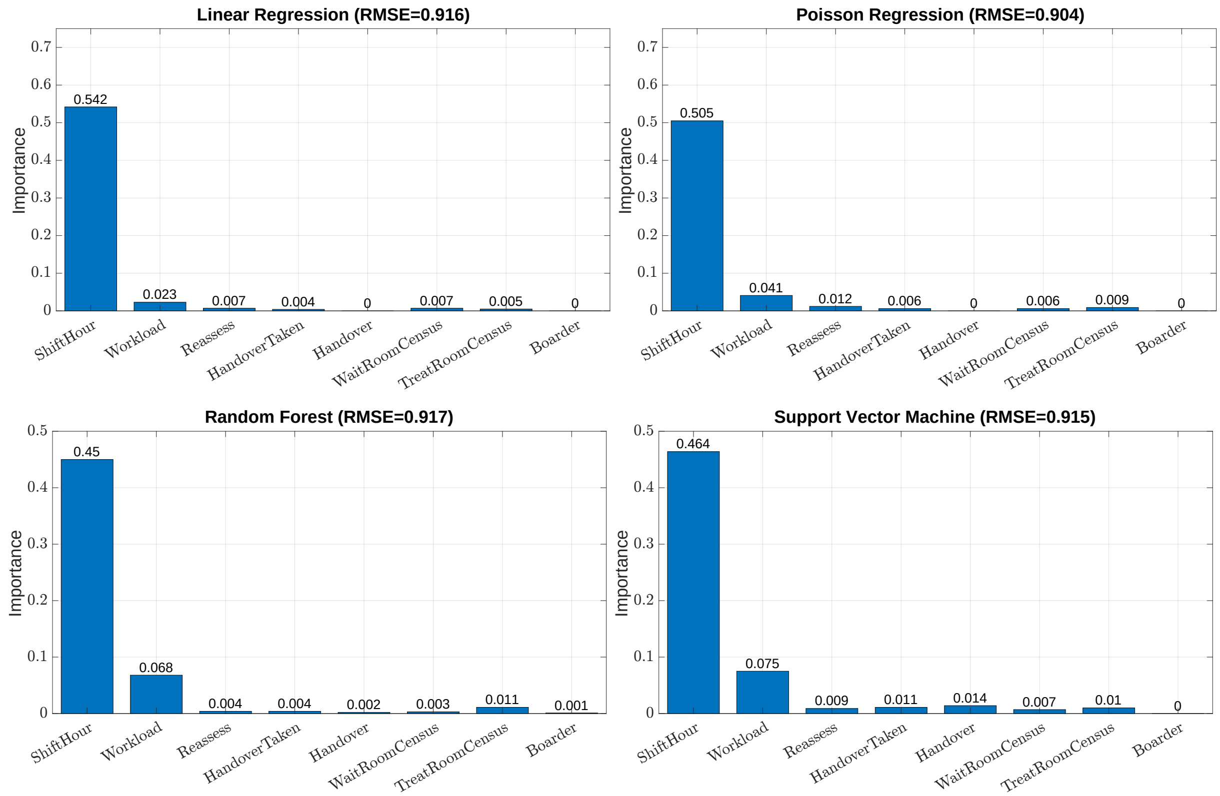
These predictive insights complement the regression findings: even in outÃ¢â‚¬â€~ofÃ¢â‚¬â€~sample prediction exercises that do not rely on parametric assumptions, intraÃ¢â‚¬â€~shift temporal dynamics remain the dominant explanatory dimension of physician productivity.

C. Discussion on Possible Mechanisms From the Literature

We believe that the observed time-varying pattern in physician productivity emerges from the aggregate effects of multiple underlying mechanisms. Following the framework of Tan and Netessine (2014), we group them into two broad categories: *mechanistic* and *behavioral*. Mechanistic mechanisms originate from structural or informational features of the work system that influence throughput without involving deliberate changes in physiciansÃ¢â‚¬â„¢ effort. In contrast, behavioral mechanisms emerge when physicians consciously or subconsciously modify their behavior, pace, or decision rules in response to workload, time pressure, or perceived incentives.

Mechanistic mechanisms. Several structural elements of ED operations can produce the observed decline. Physicians must continuously balance attention across new and existing patients. Early in a shift, the number of active cases is typically small, but as the shift progresses, prior patients remain under observation or await test results, limiting physiciansÃ¢â‚¬â„¢ capacity to take on new ones. This growing multitasking burden reduces the attention available to initiate new encounters, even without any deliberate change in effort. In

Figure A5 **The feature importance of variables using four models to predict the PPH for 8-hour shifts, and their corresponding root mean squared error (RMSE). Feature importance is calculated by permuting the corresponding variable 50 times.**



addition, frequent task switching can also negatively affect physician productivity (Duan et al. 2020). Heavy congestion in treatment areas or diagnostic facilities can further delay care steps and reduce throughput. Administrative and documentation tasks also tend to accumulate over time, diverting attention from direct patient care and lowering effective clinical capacity. Finally, structural workflow patterns, such as batching of disposition decisions near the end of shifts, can alter measured productivity without representing a true behavioral slowdown (Feizi et al. 2023). Together, these factors reflect operational constraints intrinsic to the work system rather than deliberate pacing behavior.

Behavioral mechanisms. Beyond structural constraints, physicians may adjust their behavior as they progress through the shift under different working conditions. Fatigue can gradually diminish physiciansÃ¢â‚¬â„¢ cognitive capacity and physical pace, a pattern documented among other healthcare workers (Dai et al. 2015), paramedics (Brachet et al. 2012, Bavafa and Jónasson 2024), and food inspectors (Ibanez and Toffel 2020). Near the end of a shift, the anticipated cost of handoffs often discourages physicians from initiating new cases, producing a widely observed slowdown as they reallocate effort from starting new encounters to managing existing patients, completing documentation or coordinating admissions. Familiarity and trust among physicians and nurses can promote cooperative multitasking and higher pickup rates

(Niewoehner et al. 2023), whereas lack of familiarity or uneven workload sharing may lead to coordination delays or social loafing (Berry Jaeker and Tucker 2012). When the ED becomes crowded or when waiting patients are visible on the tracking board, social accountability and peer comparison may induce a temporary speed-up, similar to the social-pressure mechanisms observed in other operational contexts (Song et al. 2015, 2018).

The time-varying PPH pattern shown in Figure 2 in the main text likely reflects the combined influence of several interacting mechanistic and behavioral mechanisms. Some factors, such as multitasking and downstream congestion, affect productivity through structural constraints, whereas others, such as pacing or social pressure, represent behavioral responses to the same set of conditions.

D. Solutions for the optimal control problem

In this section, we study the optimal control problem defined in (2) by considering three cases: (i) $R(0) = R_0 = 0, D(0) = D_0, 0 \le D_0 \le \mu_R/\theta$; (ii) $R(0) = R_0 > 0$; (iii) $R(0) = R_0 = 0, D(0) = D_0 > \mu_R/\theta$. Theorem 1, Propositions 1, 2, and 3 present the results for Case (i) since it is the most relevant case. The corresponding proofs are provided as follows. The optimal controls for Cases (ii)&(iii) are similar despite being more complicated, and their proofs are included in D.2 and D.3, respectively.

D.1. Proofs of Theorem 1 , Propositions 1, 2 , and 3 (Case (i)).

In this case, we assume $R(0) = R_0 = 0, D(0) = D_0, 0 \le D_0 \le \mu_R/\theta$. The implications of these conditions are that when the focal physician starts her shift at $t = 0$, (i) there is no patient waiting for reassessment; (ii) the number of patients who are handed off from other physicians to the focal physician and who are undergoing test is sufficiently low so that the focal physician has capacity to sign up new patients.

We first show that $R'(t) = 0,\ \forall t \in [0,T]$. When $t = 0$, because $D_0 \le \mu_R/\theta$, the constraint on $\alpha_R(t)$ in Problem (2) implies that $\alpha_R(0) = \theta D_0/\mu_R$. Hence, $R'(0) = \theta D_0 - \alpha_R \mu_R = 0$. For any $t > 0$ and $\theta D(t)/\mu_R < 1$, we have $\alpha_R(t) = \theta D(t)/\mu_R$ and hence $R'(t) = \theta D(t) - \alpha_R(t)\mu_R = \mu_R(\theta D(t)/\mu_R - \alpha_R(t)) = 0$. When $t > 0$ and $\theta D(t)/\mu_R = 1$, we have $\alpha_R(t) = 1$ and $\alpha_N(t) = 0$, which yields $R'(t) = \theta D(t) - \mu_R = 0$. Furthermore, $D'(t) = p\mu_R - \theta D(t) = -(1-p)\mu_R < 0$, which means that when $D(t)$ becomes sufficiently large such that $\theta D(t)/\mu_R = 1$, the derivative of $D(t)$ becomes negative and $D(t)$ starts to decrease. As a result, $\theta D(t+\delta t)/\mu_R < 1$ for any infinitesimally δt. Because $\theta D_0/\mu_R \le 1$, we conclude that $\theta D(t)/\mu_R \le 1$, $\alpha_R^*(t) = \theta D(t)/\mu_R$, and $R'(t) = 0,\ \forall t \in [0,T]$. Combining with $R(0) = 0$, we conclude that $R(t) = 0,\ \forall t \in [0,T]$.

Furthermore, whenever $D(t) = 0$, we have $D'(t) = p\alpha_N(t)\mu_N \ge 0$. Combining with $D(0) = D_0 \ge 0$, we conclude that the pure-state constraint $D(t) \ge 0,\ \forall t \in [0,T]$ holds naturally, which simplifies Problem (2) into the following:

$$\max_{\alpha_N(t)} \left\{ \int_0^T (1-p)\left[\alpha_N(t)\mu_N + \theta D(t)\right] \mathrm{d}t - h(D(T)) \right\} \tag{D.1}$$
$$\text{s.t.}\quad D'(t) = p\alpha_N(t)\mu_N - (1-p)\theta D(t),\ D(0) = D_0 \ge 0,\ 0 \le \alpha_N(t) \le 1 - \theta D(t)/\mu_R.$$

Next, we apply PontryaginÃ¢â‚¬â„¢s maximum principle to Problem (D.1). Denote the co-state variable of $D(t)$ by $\lambda_D(t)$. The Hamiltonian is

$$\begin{aligned} H(D,\alpha_N,\lambda_D(t),t) &= (1-p)[\alpha_N(t)\mu_N+\theta D(t)]+\lambda_D(t)\left[p\mu_N\alpha_N(t)-(1-p)\theta D(t)\right] \\ &= (1-p)\theta(1-\lambda_D(t))D(t)+\mu_N\left(p\lambda_D(t)+1-p\right)\alpha_N(t). \end{aligned} \tag{D.2}$$

Note that the Hamiltonian (D.2) is linear in $\alpha_N(t)$. The PontryaginÃ¢â‚¬â„¢s maximum principle requires the Hamiltonian to be maximized for all $t\in[0,T]$. Hence, the optimal policy to Problem (D.1) is bang-bang, i.e., $\alpha_N^*(t)$ is equal to either 0 or $1-\theta D(t)/\mu_R$. Due to the existence of the mixed inequality constraint, we need to define a Lagrangian by appending the Hamiltonian with the mixed constraints (see Chapter 3 in Sethi 2019). Let μ_L and μ_U be the Lagrange multipliers for the lower and upper constraints on the control $\alpha_N(t)$, respectively. The Lagrangian is

$$\begin{aligned} L(D,\alpha_N,\lambda_D(t),\mu_L,\mu_U,t) &= H(D,\alpha_N,\lambda_D(t),t)+\mu_L\alpha_N(t)+\mu_U\left[1-\theta D(t)/\mu_R-\alpha_N(t)\right] \\ = &\left[(1-p)\theta(1-\lambda_D(t))-\theta\mu_U/\mu_R\right]D(t)+\left[\mu_N\left(p\lambda_D(t)+1-p\right)+\mu_L-\mu_U\right]\alpha_N(t)+\mu_U. \end{aligned} \tag{D.3}$$

The optimal policy to Problem (D.1) needs to satisfy the conditions below by PontryaginÃ¢â‚¬â„¢s maximum principle:

(i) Maximum Conditions:

$$\alpha_N(t)=0 \Leftrightarrow p\lambda_D(t)+1-p<0,\ \alpha_N(t)=1-\theta D(t)/\mu_R \Leftrightarrow p\lambda_D(t)+1-p>0.$$

(ii) First-Order Conditions: $\mu_N\left(p\lambda_D(t)+1-p\right)+\mu_L-\mu_U=0.$

(iii) Complementary Slackness: $\mu_L\alpha_N(t)=\mu_U\left[1-\theta D(t)/\mu_R-\alpha_N(t)\right]=0,\ \mu_L\geq 0,\ \mu_U\geq 0.$

(iv) Adjoint Conditions: $\lambda_D'(t)=\theta\mu_U/\mu_R-(1-p)\theta(1-\lambda_D(t)),\ \lambda_D(T)=-h'(D(T)).$

Because Problem (D.1) does not contain pure-state constraints, the co-state variable $\lambda_D(t)$ is continuous in t under optimality. Consider t at which $1-\theta D(t)/\mu_R>0$ and $\alpha_N(t)=0$. Next, we prove that $\alpha_N(\hat{t})=0$ for any $\hat{t}\geq t$ under the optimal policy, which imply the optimal control $\alpha_N^*(t)$ is of threshold type.

Because of the complementary slackness and the constraint that $1-\theta D(t)/\mu_R-\alpha_N(t)>0$, we have $\mu_U=0$. The adjoint equation for $\lambda_D'(t)$ becomes

$$\lambda_D'(t)=-(1-p)\theta(1-\lambda_D(t))=(1-p)\theta\lambda_D(t)-(1-p)\theta. \tag{D.4}$$

Solving the ordinary differential equation in (D.4) yields

$$\lambda_D(t)=Ce^{(1-p)\theta t}+1, \tag{D.5}$$

where C is a constant. If $C \geq 0$, then $\lambda_D(t) \geq 1$, and hence $p\lambda_D + 1 - p = p(\lambda_D - 1) + 1 > 0$. As a result, the maximum conditions imply that $\alpha_N(t) = 1 - \theta D(t)/\mu_R$, which contradicts with $\alpha_N(t) = 0$. Hence, $C < 0$. Plugging in the expression of $\lambda_D(t)$ in (D.5) into (D.4) yields

$$\lambda_D'(t) = (1-p)\theta C e^{(1-p)\theta t} < 0,$$

whenever $\alpha_N(t) = 0$. This implies that once $\alpha_N(t) = 0$, then

$$p\lambda_D(\hat{t}) + 1 - p < p\lambda_D(t) + 1 - p < 0, \forall \hat{t} \in [t, T]$$

and thus $\alpha_N(\hat{t}) = 0$, $\forall \hat{t} \in [t, T]$. In other words, once it is optimal for the physician to stop taking new patients at t, i.e., $\alpha_N(t) = 0$, then it is optimal to stop taking new patients during the remaining time of her shift.

Assume that the physician stops taking new patients at $t^* \in [0, T]$. Then, $\alpha_N^*(t) = 1 - \theta D(t)/\mu_R$ when $t \in [0, t^*]$ and $\alpha_N^*(t) = 0$ when $t \in (t^*, T]$. The system dynamics can be described as follows:

$$\frac{\mathrm{d}D(t)}{\mathrm{d}t} = p\left(\mu_N - \frac{\mu_N}{\mu_R}\theta D(t)\right) - (1-p)\theta D(t),\ D(0) = D_0,\ t \in [0, t^*],\ \textit{and} \tag{D.6}$$

$$\frac{\mathrm{d}D(t)}{\mathrm{d}t} = -(1-p)\theta D(t),\ t \in (t^*, T]. \tag{D.7}$$

Solving the ordinary differential equations in (D.6) and (D.7) yields:

$$D(t) = \left(D_0 - \frac{p\mu_N}{\theta\left(1 - p + p\mu_N/\mu_R\right)}\right) e^{-\theta\left(1-p+p\frac{\mu_N}{\mu_R}\right)t} + \frac{p\mu_N}{\theta\left(1 - p + p\mu_N/\mu_R\right)},\ t \in [0, t^*],\ \textit{and} \tag{D.8}$$

$$D(t) = D(t^*) e^{-(1-p)\theta(t-t^*)},\ t \in (t^*, T], \tag{D.9}$$

which completes the proof for Theorem 1. □

Proof of Proposition 1. Solving (D.5) together with the boundary condition in the adjoint conditions yields

$$\lambda_D(t) = 1 - \left[1 + h'(D(T))\right] e^{(1-p)\theta(t-T)},\ t \in [0, T]. \tag{D.10}$$

Let t^* denote the optimal threshold under the optimal control policy. The maximum conditions imply that $p\lambda_D(t^*) + 1 - p = 0$. Plugging (D.10) into this equation and solving it yield

$$t^* = T - \frac{\ln[p(1 + h'(D(T)))]}{(1-p)\theta}. \tag{D.11}$$

Note that the right-hand side of (D.11) is not necessarily between 0 and T. Specifically, when $p \leq [1 + h'(D(T))]^{-1}$, we have $\ln\left[p(1 + h'(D(T)))\right] \leq 0$. Hence,

$$T - \frac{\ln[p(1 + h'(D(T)))]}{(1-p)\theta} \geq T.$$

On the other hand, when $h'(D(T) \geq e^{T(1-p)\theta}/p - 1$, we have

$$T - \frac{\ln[p(1+h'(D(T)))]}{(1-p)\theta} \leq 0.$$

Hence, we let $t^* = 0$ if the right-hand side of (D.11) is less than 0, and let $t^* = T$ if the right-hand side of (D.11) is greater than T. Hence, we get the expression of t^*. Next, we prove the monotonicity of t^* with respect to θ and p.

When $t^* = 0$ or $t^* = T$, it is obvious that t^* does not change with θ or p. Next, we consider the case $0 < t^* < T$, which implies that $p > [1 + h'(D(T))]^{-1}$. Take the derivatives of t^* with respect to θ and p, respectively, we get

$$\begin{aligned}
\frac{dt^*}{d\theta} &= \frac{\ln[p(1+h'(D(T)))]}{(1-p)\theta^2} > 0,\\
\frac{dt^*}{dp} &= -\frac{(1-p)/p + \ln[p(1+h'(D(T)))]}{(1-p)^2\theta} < 0.
\end{aligned}$$

which completes the proof for Proposition 1. □

Proof of Proposition 2. We first take the derivatives of α_N with respect to μ_N and μ_R:

$$\begin{aligned}
\frac{d\alpha_N}{d\mu_N} &= -\frac{p(1-p)\mu_R}{[p\mu_N + (1-p)\mu_R]^2}\left(1 - e^{-\theta(1-p+p\mu_N/\mu_R)t}\right) - \frac{\theta p t}{\mu_R}\left(\frac{p\mu_N\mu_R}{p\mu_N + (1-p)\mu_R} - \theta D_0\right)e^{-\theta(1-p+p\mu_N/\mu_R)t} < 0,\\
\frac{d\alpha_N}{d\mu_R} &= \frac{\theta}{\mu_R}\left(\frac{D(t)}{\mu_R} - \frac{dD(t)}{d\mu_R}\right) = \frac{D_0}{\mu_R}e^{-\theta(1-p+p\mu_N/\mu_R)t} + \frac{p(1-p)\mu_N\mu_R}{\theta[p\mu_N + (1-p)\mu_R]^2}\left(1 - e^{-\theta(1-p+p\mu_N/\mu_R)t}\right)\\
&\quad + \frac{p\mu_N t}{\mu_R^2}\left(\frac{p\mu_N\mu_R}{p\mu_N + (1-p)\mu_R} - \theta D_0\right)e^{-\theta(1-p+p\mu_N/\mu_R)t} > 0,
\end{aligned}$$

and both inequalities hold because of $D_0 \leq \frac{p\mu_N\mu_R}{\theta[p\mu_N + (1-p)\mu_R]}$. Next, taking the derivatives of α_N with respect to θ yields

$$\begin{aligned}
\frac{d\alpha_N}{d\theta} &= -\frac{D_0}{\mu_R}e^{-\theta(1-p+p\mu_N/\mu_R)t} + \left(\frac{\theta D_0}{\mu_R} - \frac{p\mu_N\mu_R}{p\mu_N + (1-p)\mu_R}\right)(1 - p + p\mu_N/\mu_R)te^{-\theta(1-p+p\mu_N/\mu_R)t}\\
&= -\frac{1}{\mu_R}e^{-\theta(1-p+p\mu_N/\mu_R)t}\left[D_0 + \frac{p\mu_N + (1-p)\mu_R}{\mu_R}\left(\frac{p\mu_N\mu_R}{p\mu_N + (1-p)\mu_R} - \theta D_0\right)t\right] < 0,
\end{aligned}$$

and the inequality holds because of $D_0 \leq \frac{p\mu_N\mu_R}{\theta[p\mu_N + (1-p)\mu_R]}$. □

Proof of Proposition 3. Because $\mathrm{PPH}(t) = \alpha_N(t)\mu_N$, we get

$$\mathrm{PPH}(t) = \frac{(1-p)\mu_N\mu_R}{p\mu_N + (1-p)\mu_R} + \mu_N\left(\frac{p\mu_N}{p\mu_N + (1-p)\mu_R} - \frac{\theta D_0}{\mu_R}\right)e^{-\theta\left(1-p+p\frac{\mu_N}{\mu_R}\right)t},\ \forall t \in [0, t^*].$$

It is obvious that $\mathrm{PPH}(t) = 0,\ \forall t \in (t^*, T]$, which completes the proof. □

D.2. The structure of the optimal policy under Case (ii).

In this case, we assume $R(0) = R_0 > 0, D(0) = D_0 \geq 0$. The implication is that when the focal physician starts her shift at $t = 0$, there are return patients waiting for reassessment. The system dynamics can be described by the ordinary differential equation as follows:

$$R'(t) = -\alpha_R \mu_R + \theta D(t),\ R(0) = R_0 > 0, \tag{D.12}$$

$$D'(t) = p\left[\alpha_N \mu_N + \alpha_R \mu_R\right] - \theta D(t),\ D(0) = D_0. \tag{D.13}$$

Since we assume that return patients are prioritized over new patients, the physician will spend 100% of her efforts to return patients, i.e., $\alpha_R(t) = 1$ and $\alpha_N(t) = 0$, until $D(t) = \mu_R/\theta$. Solving (D.12) and (D.13) jointly, we obtain the following:

$$D(t) = \left(D_0 - \frac{p\mu_R}{\theta}\right) e^{-\theta t} + \frac{p\mu_R}{\theta},\ D(t) \geq 0, \tag{D.14}$$

$$R(t) = R_0 - (1-p)\mu_R t + \left(D_0 - \frac{p\mu_R}{\theta}\right)\left(1 - e^{-\theta t}\right),\ R(t) \geq 0. \tag{D.15}$$

The next lemma shows that $R(t) = 0$ has a unique solution.

Lemma A1. *There exists $t_R > 0$ such that $R(t_R) = 0$ and $D(t_R) \leq \mu_R/\theta$.*

Proof of Lemma A1. We first re-write $R(t)$ as

$$R(t) = D_0 + R_0 - (1-p)\mu_R t - \frac{p\mu_R}{\theta} - \left(D_0 - \frac{p\mu_R}{\theta}\right) e^{-\theta t} \equiv R_1(t) - D(t), \tag{D.16}$$

where $R_1(t) = D_0 + R_0 - (1-p)\mu_R t$. It is clear that (i) $R_1(t)$ is a decreasing function in t; (ii) $R_1(0) = D_0 + R_0 > D(0) = D_0$; and (iii) $\lim_{t\to\infty} R_1(t) = -\infty < \lim_{t\to\infty} D(t) = p\mu_R/\theta$. Both $R_1(t)$ and $D(t)$ are monotonic functions of t. Hence, they have a unique intersection, i.e., $R(t) = 0$ has a unique solution.

Let t_R denote the solution to $R(t) = 0$, i.e., $R_1(t_R) = D(t_R)$. Next, we consider the following three cases separately to prove that $D(t_R) < \mu_R/\theta$: (i) $D_0 < p\mu_R/\theta$, (ii) $p\mu_R/\theta \leq D_0 < \mu_R/\theta$, and (iii) $D_0 \geq \mu_R/\theta$. When $D_0 < p\mu_R/\theta$, it is obvious that $D(t) < p\mu_R/\theta \leq \mu_R/\theta$ holds for any $t > 0$. Hence, $D(t_R) < \mu_R/\theta$. When $p\mu_R/\theta \leq D_0 < \mu_R/\theta$, it is clear that $D(t)$ decreases in t and $D(t) \leq D(0) = D_0 < \mu_R/\theta$. At last, we consider the case $D_0 \geq \mu_R/\theta$. Since $D(t)$ is a continuous function on the closed interval $[0, t_R]$ and differentiable on $(0, t_R)$, the Lagrange mean value theorem implies that there exists $t_c \in (0, t_R)$ such that

$$D'(t_c) = \frac{D(t_R) - D(0)}{t_R - 0} = \frac{D(t_R) - D_0}{t_R}. \tag{D.17}$$

Hence, we have

$$R_1'(t_R) = \frac{R_1(t_R) - R_1(0)}{t_R - 0} = \frac{D(t_R) - D_0 - R_0}{t_R} \leq D'(t_c) \leq D'(t_R), \tag{D.18}$$

where the first equality holds because $R_1(t)$ is a linear function of t, the second equality holds because $R_1(t_R) = D(t_R)$, the first inequality holds because $R_0 > 0$, and the second inequality holds because $D'(t)$ increases in t due to $D''(t) = (D_0 - p\mu_R/\theta)\theta^2 e^{-\theta t} \geq 0$. As a result, $R'(t_R) = R_1'(t_R) - D'(t_R) \leq 0$. We know from (D.14) that $R'(t) = -\mu_R + \theta D(t)$. Hence, $-\mu_R + \theta D(t_R) \leq 0$, and we have $D(t_R) \leq \mu_R/\theta$, which completes the proof. □

Lemma A1 implies that when the return patients handed off to the focal physician is exhausted at t_R, the number of patients in the test queue, i.e., $D(t_R)$, is less than μ_R/θ. Hence, the system dynamics after t_R is again the same as that under Case (i) in D.1. Therefore, the structure of the optimal policy is again of threshold type. Since the analysis is the same as that in Case (i), the details are omitted.

D.3. The structure of the optimal policy under Case (iii).

In this case, we assume $R(0) = R_0 = 0, D(0) = D_0 > \mu_R/\theta$. The implications of these conditions are that when the focal physician starts her shift at $t = 0$, (i) there is no patient waiting for reassessment; (ii) the number of patients who are handed off from other physicians to the focal physician and who are undergoing test is sufficiently high so that the focal physician has no capacity to sign up new patients.

Because $R(0) = 0$ and $D_0 > \mu_R/\theta$, we have $\alpha_R(0) = \min\{1, \theta D_0/\mu_R\} = 1$ and $\alpha_N(0) = 0$. In fact, as long as $D(t) > \mu_R/\theta$, we have $\alpha_R(t) = 1$ and $\alpha_N(t) = 0$. Hence, the system dynamics when $D(t) > \mu_R/\theta$ can be described by the ordinary differential equation as follows:

$$D'(t) = p\mu_R - \theta D(t) = \mu_R(p - \theta D(t)/\mu_R) < 0, \tag{D.19}$$

$$R'(t) = -\mu_R + \theta D(t),\ R(0) = 0, \tag{D.20}$$

which implies that $D(t)$ decreases until $D(t) = \mu_R/\theta$. Solving (D.19) together with the initial condition $D(0) = D_0$, we have

$$D(t) = \left(D_0 - \frac{p\mu_R}{\theta}\right) e^{-\theta t} + \frac{p\mu_R}{\theta}. \tag{D.21}$$

Let $D(t) = \mu_R/\theta$, we have

$$t_0 = \frac{1}{\theta} \ln \frac{\theta D_0 - p\mu_R}{(1-p)\mu_R}, \tag{D.22}$$

where t_0 is the time it takes for $D(t)$ to decrease to μ_R/θ. Because $D(t) \geq \mu_R/\theta$ for $t \in [0, t_0]$, we know from (D.20) that $R(t_0) > 0$. The dynamic of the system after t_0 will be the same as that under Case (ii) with the initial system state $D_0 = \mu_R/\theta,\ R_0 = R(t_0) > 0$. Therefore, the structure of the optimal policy is again of threshold type. Since the analysis is the same as that in Case (ii), the details are omitted.

E. Performance Evaluation Through Uniformization

In this section, we model the $M(t)/M^{\text{PPH}}(t)/s(t)$ queue by a CTMC with state jumps at discrete time epochs and apply the uniformization method (which is also referred to as *randomization* in the literature) for the performance evaluation.

We consider a daily cycle and divide the 24 hours into periods of length l, where $((j-1)l, jl]$ represents the jth period, $j = 1, \cdots, 24/l$. For staffing purposes, l is often chosen to be one hour or half an hour. We assume that the staffing level changes only at the end of each period. Let $\mathcal{S}^j$ be the set of shifts that are ongoing during the entire period j, s_j be the cardinality of $\mathcal{S}^j$, and $\mu_j(u)$ be the service rate in period j of shift $u \in \mathcal{S}^j$. We estimate $\mu_j(u)$ by the PPH in the corresponding shift hour of shift u. We further consider piece-wise constant arrival rate and let λ_j denote the arrival rate of period j. The stochastic process in period j is the same as an $M/M/s_j$ queue with heterogeneous servers, except that at the end of period j, ongoing shifts may end, and new shifts may begin, causing instantaneous transitions of system states.

Next, we model the dynamics in the jth period by a time-homogeneous CTMC. Assume that the system has been running for a sufficiently long period of time such that the probability distribution of system states at any time of day is identical for every day. Let t be the time of day and $(x(t), \mathbf{y}(t))$ be the system state at t, where $x(t)$ is the number of patients waiting to be seen, and $\mathbf{y}(t)$ is a s_j-dimensional vector whose ith element $y_i(t)$ represents the status of the physician working on the ith shift in $\mathcal{S}^j$. Specifically, $y_i(t)$ equals 0 if the physician is idling and 1 otherwise. Assume there is no unforced idling, then we have $x(t)(s_j - \sum_{i=1}^{s_j} y_i(t)) = 0$ for all t. Hence, the dimension of the state space is significantly reduced. Consider an ED with 4 physicians on duty and assume that the number of patients waiting to be seen is capped at 300. Then, the dimension of the state space is $301 + 2^4 = 317$, whereas that of the model that considers patient returns explicitly exceeds 35 million, as we discussed at the beginning of Section 5.

We apply the uniformization method to the $M/M/s_j$ queue with the uniformization constant $\Lambda_j \triangleq \lambda_j + \sum_{u \in \mathcal{S}^j} \mu_j(u)$. Let $\boldsymbol{\pi}(t)$ be the vector that represents the probability distribution of system states at t. Then, for any pair of t_1, t_2 such that $(j-1)l < t_1 < t_2 < jl$, we have

$$\boldsymbol{\pi}(t_2) = \sum_{n=0}^{\infty} p_j(n) \boldsymbol{\pi}(t_1) P_{1j}^n, \tag{E.23}$$

where $p_j(n)$ is the Poisson probability mass function with mean $(t_2 - t_1)\Lambda_j$ and P_{1j} is the transition probability matrix of the uniformized system. When $t = jl$, an instantaneous state jump will occur when there are shifts scheduled to begin or end at t. Assume that the instantaneous state transitions are governed by P_{2j}, then $\boldsymbol{\pi}(t) = \boldsymbol{\pi}(t^-)P_{2j}$, where t^- represents the time epoch just before t. Note that P_{2j} is an identity matrix if no shift begins or ends at t. We can calculate $\boldsymbol{\pi}(t)$ for any t with proper truncation of the state space and the sum of the infinite series in (E.23). With the availability of $\boldsymbol{\pi}(t)$, we can compute the long-run average ED waiting time. The waiting time calculation and the specifications of P_{1j} and P_{2j} are as follows.

We first specify the transition probability matrix P_{1j}. The transition probability of P_{1j} from $(x_1, \mathbf{y}_1)$ to $(x_2, \mathbf{y}_2)$, denoted by $p_{(x_1,\mathbf{y}_1)\to(x_2,\mathbf{y}_2)}$, is defined as follows:

$$p_{(x_1,\mathbf{y}_1)\to(x_2,\mathbf{y}_2)} = \begin{cases} \frac{\lambda_j}{\Lambda_j} & \text{if } x_1 \geq 0, x_2 = x_1+1, \mathbf{y}_1 = \mathbf{y}_2 = \mathbf{1}_{s_j}; \text{ or } x_1 = x_2 = 0, \mathbf{y}_1 \neq \mathbf{1}_{s_j}, \\ & \quad \mathbf{y}_2 = \mathbf{y}_1 + \mathbf{e}_i, \text{ where } i = \arg\max_{1\leq m\leq s_j}\{\mu_j(S_m)(1-\mathbf{y}_1^{(m)})\}, \\ \frac{\sum_{m=1}^{s_j}\mu_j(S_m)}{\Lambda_j} & \text{if } x_1 \geq 1, x_2 = x_1 - 1, \mathbf{y}_1 = \mathbf{y}_2 = \mathbf{1}_{s_j}, \\ \frac{\mu_j(S_m)}{\Lambda_j} & \text{if } x_1 = x_2 = 0, \mathbf{y}_1^{(m)} = 1, \mathbf{y}_2 = \mathbf{y}_1 - \mathbf{e}_m, 1 \leq m \leq s_j, \\ 1 - \frac{\lambda_j + \sum_{m=1}^{s_j}\mathbf{y}_1^{(m)}\mu_j(S_m)}{\Lambda_j} & \text{if } x_1 = x_2 = 0, \mathbf{y}_1 = \mathbf{y}_2 \neq \mathbf{1}_{s_j}, \\ 0 & \text{otherwise}, \end{cases}$$

where $\mathbf{e}_i$ is the ith row of the s_j-dimensional identity matrix, $\mathbf{1}_{s_j}$ is a s_j-dimensional vector with all elements equal to 1, $\mathbf{y}_1^{(m)}$ is the mth element of $\mathbf{y}_1$, and S_m is the mth shift in $\mathcal{S}^j$.

Next, we specify P_{2j} to describe the instantaneous transition of system states at the end of the jth period due to that physicians may go off-duty or begin new shifts. Let ξ_j and η_j be the numbers of physicians who go off-duty and begin new shifts at the end of period j, respectively. Note that ξ_j and η_j are known for any given schedule. If $\xi_j = \eta_j = 0$, then P_{2j} is an identity matrix of the same dimension as $\boldsymbol{\pi}^T\boldsymbol{\pi}$; otherwise, let $p_{(x_1,\mathbf{y}_1)\to(x_2,\mathbf{y}_2)}$ denote the transition probability from $(x_1, \mathbf{y}_1)$ to $(x_2, \mathbf{y}_2)$ of P_{2j}, and let Q be an $s_j \times (s_j - \xi_j)$ matrix whose column corresponds to one of the $(s_j - \xi_j)$ physicians who continue working in period $j+1$. Each column of Q is a s_j-dimensional vector whose elements are all 0 except that the gth element is 1 where g is the index of the corresponding physician in $\mathbf{y}_1$. Then, when $\eta_j \geq 1$,

$$p_{(x_1,\mathbf{y}_1)\to(x_2,\mathbf{y}_2)} = \begin{cases} 1, & \text{if } x_1 \geq \eta_j, x_2 = x_1 - \eta_j, \mathbf{y}_1 = \mathbf{1}_{s_j}, \mathbf{y}_2 = (\mathbf{y}_1 Q, \mathbf{1}_{\eta_j}); \text{ or } x_1 \leq \eta_j - 1, x_2 = 0, \\ & \quad \mathbf{y}_2 = (\mathbf{y}_1 Q, \mathbf{1}_{x_1}, \mathbf{0}_{\eta_j - x_1}); \text{ or } x_1 = x_2 = 0, \mathbf{y}_2 = (\mathbf{y}_1 Q, \mathbf{0}_{\eta_j}), \\ 0, & \text{otherwise}, \end{cases}$$

where $\mathbf{1}_g$ and $\mathbf{0}_g$ are g-dimensional vectors with all elements equal to 1 and 0, respectively. Note that when $x_1 < \eta_j$, i.e., the number of waiting patients is less than the number of physicians who begin their shifts, we assign the patients in waiting to newly-arrived physicians with the highest service rates, following the same rule described in Section 5.1. This is achieved by ranking the newly-arrived physicians by their current service rates. Specifically, in $\mathbf{y}_2 = (\mathbf{y}_1 Q, \mathbf{1}_{x_1}, \mathbf{0}_{\eta_j - x_1})$, $\mathbf{1}_{x_1}$ represents the x_1 physicians with the highest service rates among the η_j newly-arrived physicians and $\mathbf{0}_{\eta_j - x_1}$ represents the remaining $\eta_j - x_1$ slower physicians. When $\eta_j = 0$, then

$$p_{(x_1,\mathbf{y}_1)\to(x_2,\mathbf{y}_2)} = \begin{cases} 1, & \text{if } x_1 = x_2, \mathbf{y}_2 = \mathbf{y}_1 Q, \\ 0, & \text{otherwise}, \end{cases}$$

Finally, we follow Liu and Xie (2021) to compute the average patient waiting time. The total expected patient waiting time in the jth period $((j-1)l, jl]$, denoted by W_j, can be expressed as

$$W_j = \sum_{n=0}^{\infty}\left(\frac{(\Lambda_j l)^n}{n!}e^{-\Lambda_j l}\sum_{m=0}^{n}\sum_{i=1}^{\infty}\boldsymbol{\pi}_i((j-1)l + ml/(n+1))\frac{il}{n+1}\right),$$

where $\boldsymbol{\pi}_i((j-1)l + ml/(n+1))$ is the probability at state $(i, \mathbf{1}_{s_j})$ at $t = (j-1)l + ml/(n+1)$. Hence, the average ED waiting time is $\sum_{j=1}^{24/l} W_j / \sum_{j=1}^{24/l} \lambda_j$. Interested readers are referred to the appendix of Liu and Xie (2021) for detailed derivation and explanation.

References

H. Bavafa and J. O. Jónasson. The distributional impact of fatigue on performance. *Management Science*, 70(5): 3319–3337, 2024.

J. A. Berry Jaeker and A. L. Tucker. Hurry up and wait: Differential impacts of congestion, bottleneck pressure, and predictability on patient length of stay. *Harvard Business School working paper series# 13-052*, 2012.

T. Brachet, G. David, and A. M. Drechsler. The effect of shift structure on performance. *American Economic Journal: Applied Economics*, 4(2):219–246, 2012.

H. Dai, K. L. Milkman, D. A. Hofmann, and B. R. Staats. The impact of time at work and time off from work on rule compliance: the case of hand hygiene in health care. *Journal of Applied Psychology*, 100(3):846, 2015.

Y. Duan, Y. Jin, Y. Ding, M. Nagarajan, and G. Hunte. The cost of task switching: Evidence from the emergency department. *Available at SSRN 3756677*, 2020.

A. Feizi, A. Carson, J. B. Jaeker, and W. E. Baker. To batch or not to batch? impact of admission batching on emergency department boarding time and physician productivity. *Operations Research*, 71(3):939–957, 2023.

A. Fisher, C. Rudin, and F. Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177): 1–81, 2019.

M. R. Ibanez and M. W. Toffel. How scheduling can bias quality assessment: Evidence from food-safety inspections. *Management Science*, 66(6):2396–2416, 2020.

R. Liu and X. Xie. Weekly scheduling of emergency department physicians to cope with time-varying demand. *IISE Transactions*, 53(10):1109–1123, 2021.

R. J. Niewoehner, K. Diwas, and B. Staats. Physician discretion and patient pick-up: How familiarity encourages multitasking in the emergency department. *Operations Research*, 71(3):958–978, 2023.

S. Sethi. *Optimal Control Theory—Applications to Management Science and Economics*. Springer, third edition, 2019.

H. Song, A. L. Tucker, and K. L. Murrell. The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science*, 61(12):3032–3053, 2015.

H. Song, A. L. Tucker, K. L. Murrell, and D. R. Vinson. Closing the productivity gap: Improving worker productivity through public relative performance feedback and validation of best practices. *Management Science*, 64(6): 2628–2649, 2018.

T. F. Tan and S. Netessine. When does the devil make work? an empirical study of the impact of workload on worker productivity. *Management Science*, 60(6):1574–1593, 2014.