Appendix A (for Online Publication)

A-1 Model Setup

In this section, I outline an inclusive setup of the model that spans several assumptions, in both the nurse-managed system and the self-managed system. For detailed comments of the setup for the baseline model, refer to Section 3 in the main paper. Consider the following simple game, involving two physicians $j \in \{1, 2\}$ who work in a single pod at the same time:

1. The triage nurse (in the nurse-managed system) or physicians (in the self-managed system) commit to an assignment policy function if they are able to.

2. At time $t = 0$ both physicians receive one patient each, discovering $\theta_j \in \left\{\underline{\theta}, \overline{\theta}\right\}$, where $\overline{\theta} > \underline{\theta} > 0$. Type $\underline{\theta}$ occurs with probability p, and type $\overline{\theta}$ occur with probability $1 - p$.[1]

3. Physicians commit to how long they will keep their initial patients (t_j).

4. Physicians send a message $m_j \in \left\{\underline{\theta}, \overline{\theta}\right\}$ to the triage nurse (in the nurse-managed system) or to each other (in the self-managed system) if they are able to report their types. Otherwise they send $m_j = \emptyset$.

5. With probability $\psi > 0$, physicians observe each other's θ_j. Physician types are never observed by the triage nurse.

6. Exactly one patient will arrive at time t_a distributed with uniform probability across $\left[\underline{\theta}, \overline{\theta}\right]$. The physician who receives this third patient, denoted by$\mathcal{J}(3)$, is determined as follows:

   (a) In the nurse-managed system, the triage nurse assigns the patient based on physician censuses c_1 and c_2 (the number of patients they have, either 0 or 1, which is public information) at $t = t_a$. If she committed to an assignment policy function in Stage 1, she uses this. Otherwise, she decides assignment at t_a. If physicians can report their types, then her policy function can also use m_1 and m_2.

   (b) In the self-managed system, the physicians determine assignment.

      i. If physicians cannot commit to an assignment policy function, each physician independently decides whether to choose the new patient at each time $t \geq t_a$.

      ii. If physicians can commit to an assignment policy function, then the new patient is assigned according to this policy function at $t = t_a$. The policy function uses censuses c_1 and c_2, and observed types $o_j \in \{\theta_j, \emptyset\}$ ($o_j = \theta_j$ with probability ψ, $o_j = \emptyset$ with probability $1 - \psi$). If physicians can report their types, then the policy function also uses m_j.

[1] In this appendix, I will refer to θ_j as the type of physician j, since physicians start with one patient each and are otherwise identical.

7. Physicians complete their work on the one or two patients under their care and end their shifts. They receive the payoff

$$u_j^P = -\left(t_j - \theta_j\right)^2 - K_P\left(\theta_j\right)\mathbb{I}\left\{\mathcal{J}\left(3\right) = j\right\}, \tag{A-1.1}$$

where t_j is the time that physician j keeps his initial patient, $\theta_j \in \left\{\underline{\theta}, \overline{\theta}\right\}$ is the workload entailed by his patient and unobservable by the triage nurse, and $K_P\left(\theta_j\right) > 0$ is the cost of getting a potential third patient conditional on initial workload θ_j. I denote $K_P\left(\underline{\theta}\right) = \underline{K}_P$ and $K_P\left(\overline{\theta}\right) = \overline{K}_P$, and I impose that $\overline{K}_P > \underline{K}_P > 0$.

A-2 Nurse-managed System

In the nurse-managed system, the triage nurse assigns the new patient to a physician. I flexibly specify the triage nurse's utility as

$$u^N = -D \sum_{j \in \{1,2\}} \left(t_j - \theta_j\right)^2 - K_N\left(\theta_{j(3)}\right). \tag{A-2.1}$$

Notice the similarity between this utility function and that of the physicians, shown in (A-1.1). D is an indicator that allows the triage nurse to care about the treatment times of the first two patients as outcomes (if $D = 1$). Remember that the socially optimal discharge times for patients is $t_j = \theta_j$, which is universally agreed upon. The second term, $K_N\left(\theta\right)$, is the cost of assigning the new patient to a physician of type θ. I specify $K_N\left(\underline{\theta}\right) = 0$ and $K_N\left(\overline{\theta}\right) = \overline{K}_N$, and I impose $\overline{K}_N > 0$ but do not restrict the the value of $\overline{K}_N$ relative to $\overline{K}_P - \underline{K}_P$. This reflects that the triage nurse would like to assign the new patient to a physician with low workload.

The triage nurse's action is defined by an assignment policy function $\pi\left(c_1, c_2\right)$, where censuses $c_j \in \{0, 1\}$ are the numbers of patients the physicians have at time t_a.[2] To simplify notation, I impose that $\pi\left(0,0\right) = \pi\left(1,1\right) = \frac{1}{2}$ and $\pi \equiv \pi\left(0,1\right) = 1 - \pi\left(1,0\right)$. That is, when both physicians have equal censuses, assignment should not prefer to send the new patient to one physician or the other, since there is no other information about who is less busy at that time. Also, probabilities must sum to 1. To be clear, $\pi = 1$ represents *ex post* efficiency in that if one physician has no patients and the other has one, the former physician is known to be less busy with certainty.

In what follows, I will first show that the analysis is simplified by the fact that each physician's best response function is unaffected by what he expects the other physician to do. I then consider three different scenarios of whether physicians can report their types and of whether the triage nurse can commit to an assignment policy function. In my preferred model, I assume that physicians cannot report their types but that the triage nurse can credibly commit to an assignment policy. Although θ_j is a single index known with certainty in this model, in practice

[2] Below, I consider the scenario in which physicians can report their types to the triage nurse. In this case, the policy function takes the form $\pi\left(m_1, m_2\right)$, where $m_j \in \left\{\underline{\theta}, \overline{\theta}\right\}$.

workload is difficult to communicate, and ED physicians rarely report their workloads to the triage nurse.[3] On the other hand, what the triage nurse does when she observes c_1 and c_2 is easily observable; the ED management may even set guidelines which instruct her to assign not entirely by censuses, and more importantly, beliefs about her behavior can easily be updated in practice by physicians when there is more than one new patient arriving during their shifts. Finally, I slightly modify the model to communicate the intuition for identification strategy of using the expected flow of future patients as a driver of foot-dragging.

A-2.1 Irrelevance of Peer Strategy or Type

One advantage for the uniqueness of equilibria in this simple two-type model is that the physician's strategy does not depend on the strategy or type of his peer. To see this, consider some assignment function $\pi(c_1, c_2)$, which we will assume for now is based on censuses. Again, this assignment function can be summarized by a single parameter $\pi \equiv \pi(0,1)$. Any reasonable assignment rule sends the new patient to a less-busy physician, if there is one, with greater probability (i.e., $\pi > \frac{1}{2}$). It is easy to see that a high-type physician will then discharge his patient at $\overline{t} = \overline{\theta}$, because discharging his patient earlier only increases his changes of getting the new patient and discharging later has no benefit either.

Lemma A-1. *A high-type physician always discharges his patient at $\overline{t} = \overline{\theta}$.*

The next step is to solve for the best response function of the index physician who is of low type, given his peer's strategy. Denote the discharge time that a low-type peer will choose as $\underline{t}_{-j} \in \left[\underline{\theta}, \overline{\theta}\right]$. The expected utility of the index physician is

$$\mathbb{E}\left[u_j^P\left(\underline{t}_j; \pi, \underline{\theta}\right)\right] = -\left(\underline{t}_j - \underline{\theta}\right)^2 - \underline{K}_P \Pr\left\{\mathcal{J}(3) = j \left|\underline{t}_j, \pi\right.\right\},$$

The probability that $\mathcal{J}(3) = j$ can be written more explicitly as

$$\Pr\left\{\mathcal{J}(3) = j \left|\underline{t}_j, \pi\right.\right\} = \begin{cases} \frac{1}{2}\left(\frac{\underline{t}_j - \underline{\theta}}{\overline{\theta} - \underline{\theta}}\right) + \pi\left(\frac{\underline{t}_{-j} - \underline{t}_j}{\overline{\theta} - \underline{\theta}}\right) + \left(\frac{1}{2}p + \pi(1-p)\right)\left(\frac{\overline{\theta} - \underline{t}_{-j}}{\overline{\theta} - \underline{\theta}}\right), & \underline{t}_j < \underline{t}_{-j} \\ \begin{array}{c} \frac{1}{2}\left(\frac{\underline{t}_{-j} - \underline{\theta}}{\overline{\theta} - \underline{\theta}}\right) + \left((1-\pi)p + \frac{1}{2}(1-p)\right)\left(\frac{\underline{t}_j - \underline{t}_{-j}}{\overline{\theta} - \underline{\theta}}\right) + \\ \left(\frac{1}{2}p + \pi(1-p)\right)\left(\frac{\overline{\theta} - \underline{t}_j}{\overline{\theta} - \underline{\theta}}\right) \end{array}, & \underline{t}_j > \underline{t}_{-j} \end{cases}, \tag{A-2.2}$$

which is continuous at $\underline{t}_j = \underline{t}_{-j}$. This expression represents flow probabilities divided among three potential windows of time and uses the fact that the peer is low-type with probability p.

[3] Patient status in the ED is highly uncertain and multidimensional (e.g., patients expected to have the same length of stay may entail very different workloads due to severity). Physicians often do not share the same assessment when viewing the same patient, and information is invariably lost in communication, such that patient "hand-offs" are viewed as a patient safety issue (Apker et al., 2007). For similar reasons, it is highly unlikely that physicians may have a credible way report each other's workloads to the triage nurse, as conceived by Moore and Repullo (1988) as a subgame perfect mechanism. This is discussed more at the beginning of Section A-2.4.

Both $\underline{t}_{-j}$ and p additively affect expected utility, but they do not affect the first-order condition with respect to $\underline{t}_j$. In particular,

$$\frac{\partial}{\partial \underline{t}_j} \Pr\left\{\mathcal{J}(3) = j \left| \underline{t}_j, \pi \right.\right\} = \frac{1}{\overline{\theta} - \underline{\theta}} \left(\pi - \frac{1}{2}\right). \tag{A-2.3}$$

Increasing $\underline{t}_j$ always reduces the flow probability of receiving the third patient by $\pi - \frac{1}{2}$. Thus, his best response does not depend on the action of his (low-type) peer or the probability that his peer is low-type. It consequently is immaterial whether he knows his peer's type or the sequence in which physicians first choose t_j and then observe θ_{-j} with probability ψ.

Lemma A-2. *The best response of a low-type physician in the nurse-managed system does not depend on the action of his peer nor the probability that his peer is low-type.*

A-2.2 No Physician Reporting, No Triage Nurse Commitment

Without physician reporting or triage nurse commitment, in equilibrium, the triage nurse chooses the optimal assignment policy $\pi(c_1, c_2)$ at time t_a, given physician discharge strategies $\underline{t}^*$ and $\overline{t}^*$ for low- and high-type physicians, respectively, and given censuses c_1 and c_2. With no commitment, the triage nurse will always choose $\pi^* \equiv \pi^*(0, 1) = 1$, the *ex post* efficient choice, and physicians will respond with discharge strategies $\underline{t}^*$ and $\overline{t}^*$.

Proposition A-3. *In the Perfect Bayesian Equilibrium in the nurse-managed system with no physician reporting and no triage nurse commitment, the triage nurse always assigns the new patient to the physician with census 0 when there is another physician with census 1, i.e.,* $\pi^* = 1$. *Low-type physicians will foot-drag, choosing* $\underline{t}^* > \underline{\theta}$, *as in Equation (A-2.4).*

The triage nurse's assignment policy is simple: She will assign the new patient to the physician with no patients if the other has one patient (i.e., $\pi^* = 1$). In this case, given any $\underline{t}^*$ and $\overline{t}^*$, she expects that the physician with $c_j = 0$ must be low-type whereas the one with $c_{-j} = 1$ must be high type. Otherwise, if $c_1 = c_2$, she cannot distinguish the two physicians and will assign the patient with equal probability to each.

In equilibrium, physicians will discharge their initial patients with this knowledge. As stated in Lemma A-1, high-type physicians will never want to mimic low-type physicians and will discharge their patients at time $t = \overline{\theta}$. On the other hand, low-type physicians will want to mimic high-type physicians at least temporarily by keeping their patients longer than socially optimal, since this reduces his likelihood of getting the new patient. (Low-type) physicians do not consider the type of their peer because of Lemma A-2. The first-order condition of a low-type physician's problem of $\max_{t_j} \mathbb{E}\left[u_j^P(t_j; \underline{\theta})\right]$ yields

$$\underline{t}^* = \underline{\theta} + \frac{K_P}{4\left(\overline{\theta} - \underline{\theta}\right)}. \tag{A-2.4}$$

This reflects the fact that there is always a first-order gain to reducing the likelihood of getting the new patient, compared to a second-order loss to prolonging the patient stay to more than socially optimal. That is, there will always be foot-dragging. In fact, for sufficiently high $\underline{K}_P$ or sufficiently low $\overline{\theta} - \underline{\theta}$, there may be full pooling in that all physicians will discharge their patient at $t = \overline{\theta}$.

In this subgame perfect equilibrium, the triage nurse correctly assigns patients to physicians who are less busy when she observes $c_1 \neq c_2$. But by doing so, she incentivizes physicians to foot-drag and actually reduces the probability of seeing $c_1 \neq c_2$ when $\theta_1 \neq \theta_2$. That is, while her assignment is *ex post* efficient, it is *ex ante* inefficient.

A-2.3 No Physician Reporting, Triage Nurse Commitment

I now allow for the triage nurse to commit to a policy function π_N. The equilibrium in this case is the same as in the previous case without commitment, except that the triage nurse chooses π_N^* at $t = 0$ and not at $t = t_a$.

Proposition A-4. *In the Perfect Bayesian Equilibrium in the nurse-managed system with no physician reporting but triage nurse commitment, the triage nurse assigns the new patient to the physician with census 0 when there is another physician with census 1 with some probability π_N^* as given by Equation (A-2.7), which lies between $\frac{1}{2}$ and 1. A low-type physician foot-drags weakly less than under no triage nurse commitment, discharging his patient at $\underline{t}^*$ that is earlier and closer to $\underline{\theta}$.*

To analyze this case, first note that the triage nurse will still never want to send the new patient with greater probability to a physician with $c_j > c_{-j}$. It therefore still holds that high-type physicians will never want to mimic low-type physicians, and that low-type physicians have some reason to mimic high-type ones. To see this, for any given π_N, the first-order condition for a low-type physician yields

$$\underline{t}^* = \underline{\theta} + \frac{\underline{K}_P}{2\left(\overline{\theta} - \underline{\theta}\right)} \left(\pi_N - \frac{1}{2}\right). \tag{A-2.5}$$

The first-order gain in temporary mimicry still exists relative to the second-order loss, as long as $\pi_N > \frac{1}{2}$.

When the triage nurse commits to an assignment policy function, she will choose π_N such that her expected utility is maximized. Her expected utility is

$$\begin{aligned}\mathbb{E}\left[u^N\left(\pi_N\right)\right] \quad = \quad & -Dp\left(\underline{t}^* - \underline{\theta}\right)^2 - (1-p)^2 \overline{K}_N - \\ & 2p\left(1-p\right)\overline{K}_N \left[\frac{1}{2}\left(\frac{\underline{t}^* - \underline{\theta}}{\overline{\theta} - \underline{\theta}}\right) + \left(1 - \pi_N\right)\left(\frac{\overline{\theta} - \underline{t}^*}{\overline{\theta} - \underline{\theta}}\right)\right].\end{aligned} \tag{A-2.6}$$

Substituting (A-2.5) into (A-2.6) and solving the first-order condition yields the optimal assignment rule under commitment

$$\pi_N^* = \frac{1}{2} + \frac{4p\left(1-p\right)\overline{K}_N\left(\overline{\theta}-\underline{\theta}\right)^2}{\underline{K}_P\left(Dp\underline{K}_P + 4p\left(1-p\right)\overline{K}_N\right)}. \tag{A-2.7}$$

It is easy to see that π_N^* given by (A-2.7) can be less than 1, which implies that the triage nurse sometimes sends the new patient to the physician with census 0 even when there is another one with census 1. With $\pi_N^* < 1$, by (A-2.5), $\underline{t}^*$ is lower than when there is no triage nurse commitment.

The important general point is that for some parameters the triage nurse will commit to an assignment policy function $\pi_N^* < 1$. Even if she only cares about the assignment of the third patient, as in Equation (A-2.8), she will commit to an *ex post* inefficient assignment policy in order to improve *ex ante* assignment. Under triage nurse commitment to $\pi_N^* < 1$, the degree of foot-dragging by low-type physicians, in Equation (A-2.5), will be lower than under no commitment, stated in Equation (A-2.4). The intuition for this is similar to the finding in Milgrom and Roberts (1988) that managers will sometimes choose to ignore valuable but distortable information from "influence activities."

This result is even stronger when the triage nurse cares about length of stay for the initial patients (i.e., $D = 1$). To see this, consider the optimal policy function if the triage nurse does not care about lengths of stay for the first two patients as outcomes and only cares about the assignment of the third patient (i.e., $D = 0$):

$$\pi_{N|D=0}^* = \frac{1}{2} + \frac{\left(\overline{\theta}-\underline{\theta}\right)^2}{\underline{K}_P}. \tag{A-2.8}$$

The triage nurse's choice of $\pi_{N|D=0}^*$ only depends on the *low-type physician's* cost of getting the new patient, because it is this physician that will engage in foot-dragging and distort information. In the case with $D = 1$, in (A-2.7), $\pi_{N|D=1}^*$ also increases as $\overline{K}_N$ or as p is closer to $\frac{1}{2}$. Since she cares about foot-dragging as an outcome, she will increase $\pi_{N|D=1}^*$ as she cares more about inefficient assignment (as $\overline{K}_N$ is higher) or as it is more likely she will encounter two physicians with different censuses (p is closer to $\frac{1}{2}$). Importantly, comparing (A-2.8) and (A-2.7), $\pi_{N|D=0}^* > \pi_{N|D=1}^*$. Also caring about foot-dragging as an outcome lowers her choice of π_N^* to reduce foot-dragging further.

A-2.4 Physician Reporting, Triage Nurse Commitment

I next consider the case in which physicians can also report their types as $m_j \in \left\{\underline{\theta}, \overline{\theta}\right\}$, which is a mechanism design problem without transfers. In the equilibrium in this case, the triage nurse provides a menu $\left\{t\left(\underline{\theta}\right), t\left(\overline{\theta}\right); \pi\left(m_1, m_2\right)\right\}$ to physicians, subject to an incentive compatibility constraint for physicians to report $m_j = \theta_j$ truthfully. Because the triage nurse immediately distributes the new patient upon arrival, there is an additional "intertemporal" constraint in that a physician who lies and then later reveals that he is lying can still be punished sufficiently so that he will not want to engage in this strategy. Given this menu, physicians will tell the truth.[4]

I find that, in contrast to physician signaling (no reporting to the triage nurse), there is no foot-dragging under the mechanism of physician reporting when there are discrete types. However, this is not the case with continuous types, in which physicians always have a first-order incentive to foot-drag. I show results for the case of continuous types in Section A-4, but the intuition is consistent with that of Myerson and Satterthwaite (1983): Efficiency is improved somewhat artificially when one restricts the types (and messages) that physicians can report. The restriction of the type space is quite severe in the ED (and most medical contexts), as patient care and physician workload are multidimensional and contain quite a lot of information, to the extent that communication problems exist even between doctors with no incentive for moral hazard (e.g., Apker et al., 2007). Multidimensional screening is even more complex than screening along a single continuous dimension and has been discussed in Rochet and Stole (2003).

Proposition A-5. *In the Perfect Bayesian Equilibrium in the nurse-managed system with physician reporting and triage nurse commitment, the triage nurse implements truth-telling by setting the assignment policy π_R^* equal to the assignment policy implied by intertemporal incentive compatibility, π_{IT}^*, given in Equation (A-2.12). Given truth-telling and discrete types, there is no foot-dragging, i.e., $\underline{t}^* = \underline{\theta}$. Assignment is still* ex post *inefficient with $\pi_R^* < 1$, but* ex ante *assignment efficiency is improved relative to no physician reporting.*

By Lemma A-1, there is no incentive compatibility constraint for high-type physicians, but there is one for low-type physicians. Specifically, the utility for a low-type physician under truth-telling must be at least as great as his utility if he were to report that he is a high-type physician, where I again use the notation $\pi \equiv \pi\left(\underline{\theta}, \overline{\theta}\right)$ to summarize any assignment policy:

$$-\left(t\left(\overline{\theta}\right) - \underline{\theta}\right)^2 - \frac{1}{2}\underline{K}_P \leq -\pi_{IC}\underline{K}_P. \tag{A-2.9}$$

Using the fact that $t\left(\overline{\theta}\right) = \overline{\theta}$ by Lemma A-1,[5] this incentive compatibility constraint is actually binding at the optimal triage nurse assignment when she only cares about foot-dragging as a signal, stated in Equation (A-2.8):

$$\pi_{IC}^* = \frac{1}{2} + \frac{\left(\overline{\theta} - \underline{\theta}\right)^2}{\underline{K}_P}. \tag{A-2.10}$$

Note that although π_{IC}^* equals $\pi_{N|D=0}^*$ in Equation (A-2.8), the intuitions for the two expressions are different. In the case of π_{IC}^*, the constraint sets utility to be the same for a low-type physician

[4] Again, expectations over the peer's type are conveniently ignored due to Lemma A-2.

[5] In some mechanism design problems, this value of a high-type agent could be distorted upwards. However, because I have assumed that there is 0 probability that the new patient will arrive after $t = \overline{\theta}$, this mechanism cannot be enforced.

under truth-telling and under reporting to be high-type. In the case of $\pi^*_{N|D=0}$, triage nurse utility happens to be maximized at an assignment policy that causes a low-type physician to foot-drag midway with $\underline{t}^* = \underline{\theta} + \left(\overline{\theta} - \underline{\theta}\right)/2$. Also recall that $\pi^*_{N|D=0} > \pi^*_{N|D=1}$; so $\pi^*_{IC} \leq \pi^*_N$.

However, there is a second "intertemporal" incentive compatibility constraint that results from the facts that reports of θ_j and discharge times t_j are intertemporally separated and that the triage nurse is limited in her ability to punish a physician who reports $\theta_j = \overline{\theta}$ but discharges his patients before $t\left(\overline{\theta}\right) = \overline{\theta}$. That is, a low-type physician may report that he is high-type but discharge his patient at $\hat{t} < \overline{\theta}$. Without punishment (i.e., if he simply reverts to having the truth-telling flow probability of $\pi_{IT}\left(\underline{\theta}, \overline{\theta}\right)$ after t) he would be strictly better off by this scheme. To prevent this, there needs to be punishment, but the highest flow probability that the triage nurse can use from that point on is 1.

The second constraint is thus

$$-\left(\hat{t} - \underline{\theta}\right)^2 - \underline{K}_P\left[\frac{1}{2}\left(\frac{\hat{t} - \underline{\theta}}{\overline{\theta} - \underline{\theta}}\right) + \frac{\overline{\theta} - \hat{t}}{\overline{\theta} - \underline{\theta}}\right] \leq -\pi_{IT}\underline{K}_P. \tag{A-2.11}$$

This can be addressed by first solving for the optimal "cheating" $\hat{t}^*$ under full punishment:

$$\hat{t}^* = \underline{\theta} + \frac{\underline{K}_P}{4\left(\overline{\theta} - \underline{\theta}\right)},$$

which of course is the same as the optimal discharge time with no physician reporting or triage nurse commitment, stated in (A-2.4). This time can then be substituted into (A-2.11) in order to state the constraint on π_{IT}:

$$\pi^*_{IT} = 1 - \frac{\underline{K}_P}{16\left(\overline{\theta} - \underline{\theta}\right)^2}. \tag{A-2.12}$$

Note that $\pi^*_{IT} < 1$ for $\underline{K}_P > 0$. The intutition for this is that, with intertemporal incentive compatibility, the triage nurse can never implement $\pi_R = 1$, because if she did, then a low-type physician will always be better off by reporting to be high-type for some time. Revealing that he lied entails "punishment" which cannot be worse than the $\pi_R = 1$ he would have gotten with truth-telling anyway.[6]

It can be shown that $\pi^*_{IC} \leq \pi^*_{IT}$. In particular, $\pi^*_{IC} < \pi^*_{IT}$, for $\underline{K}^P / \left(\overline{\theta} - \underline{\theta}\right)^2 \in (0, 4)$, when there is a temptation to foot-drag but when $\underline{t}^* < \overline{\theta}$ in the case without physician reporting or triage nurse commitment, as in Section A-2.3 and Equation (A-2.4). The intuition is that it is weakly more difficult to implement the intertemporal constraint than the standard incentive compatibility constraint because it is always at least as easy for low-type physicians to "temporarily"

[6] In contrast, $\pi^*_{IC} = 1$ for $\underline{K}^P / \left(\overline{\theta} - \underline{\theta}\right)^2 \leq 2$. In this parameter space, a low-type physician is better off by telling the truth than by "fully" lying by reporting that he is high-type and keeping his patient until $t\left(\overline{\theta}\right) = \overline{\theta}$. Similarly, \pi_{N}^{*} = 1 for $\underline{K}^P / \left(\overline{\theta} - \underline{\theta}\right)^2 \leq 2$, because the temptation to foot-drag is sufficiently low so that the triage nurse is better off by incurring the maximum foot-dragging and the highest *ex post* assignment efficiency.

lie rather than “fully” lie. Thus, the intertemporal constraint will always be binding.

The assignment policy under physician reporting is therefore $\pi_R^* = \min(\pi_{IC}^*, \pi_{IT}^*) = \pi_{IT}^*$. Again, the assignment policy under no physician reporting but triage nurse commitment is equivalent to the assignment policy implied by the constraint with “full” lying: $\pi_{N|D=0}^* = \pi_{IC}^*$. At first glance, this suggests that assignment is less efficient *ex post* under physician reporting than under no reporting: $\pi_{N|D=0}^* < \pi_R^*$. This may not hold if the triage nurse also cares about foot-dragging on the initial patients ($D = 1$) under no reporting, since $\pi_{N|D=0}^* > \pi_{N|D=1}^*$.

However, the more important point is that under reporting and truth-telling, *ex post* assignment efficiency is also *ex ante* assignment efficiency, since physicians do not foot-drag. Further, the triage nurse's *ex ante* utility is strictly greater with physician reporting than with no reporting. To formalize this, I consider $\mathbb{E}\left[u^N\right]$, where $u^N$is given in Equation (A-2.1), which simplifies to $\overline{K}^N \Pr\left\{\theta_{j(3)} = \overline{\theta}\right\}$ when $D = 0$. Figure A-2.2 shows the difference in this value between physician reporting and no reporting, normalizing $\overline{K}^N = 1$. Note that this efficiency gain is strictly greater when $D = 1$, because the triage nurse also cares about foot-dragging on the initial patients as negative outcomes, of which there is none under reporting.

In summary, Proposition A-5 derives from the triage nurse's access to better information from physician reports. Subject to maintaining truth-telling, she can implement an *ex ante* assignment policy more efficient than the one without physician reporting, when she must balance *ex post* assignment efficiency with distortion of signals by physicians. In both cases, she is limited by the ability of physicians to distort signals or misreport the truth. In this sense, there is also a parallel intuition between the Milgrom and Roberts (1988) prediction and the standard mechanism design feature of information rents.

In this two-type model, there is no foot-dragging because the triage nurse uses π_R^* in order to implement truth-telling. There is no point in using $t(\underline{\theta})$ to implement truth-telling, because doing so would only make a low-type physician worse off under truth-telling, and $t\left(\overline{\theta}\right)$ is irrelevant because of the intertemporal incentive compatibility constraint. However, I will show in Section A-4 that this does not hold for continuous types, because with local incentive compatibility constraints, the benefit of foot-dragging (“full” lying in the mechanism design framework) at the truth ($t_j = \theta_j$) is first-order while its cost is second-order. As argued above, this is an important concern for the ED setting (and most other medical settings) in which information is quite complex, and in which the restriction to discrete types is highly artificial.

A-2.5 Flow of Expected Future Work

This simple model assumes a single patient will arrive in the interval $t \in \left[\underline{\theta}, \overline{\theta}\right]$, which is convenient for capturing the temptation for moral hazard by low-type physicians. However, there are of course in practice usually many more new patients, and my main identification for foot-dragging will be the response of lengths of stay to the flow of expected future work. One way to capture the intuition of expected future work is to modify the model so that a new patient is expected to arrive in the interval $t \in [\underline{\theta}, \underline{\theta} + \Delta t]$, maintaining the assumption of a single new patient. I also allow Δt to be greater or less than $\overline{\theta} - \underline{\theta}$, although I assume that physicians may only be assigned the new patient prior to $t = \overline{\theta}$ for $\Delta t > \overline{\theta} - \underline{\theta}$ and focus on interior solutions for $\Delta t < \overline{\theta} - \underline{\theta}$.[7]

I again focus on the behavior of low-type physicians, who has an incentive to foot-drag and mimic high-type physicians. It is easy to see that all denominators in fractions in Equation (A-2.2) should now be Δt instead of $\overline{\theta} - \underline{\theta}$, and $\overline{\theta}$ should be replaced by $\underline{\theta} + \Delta t$. Equation (A-2.3) should then have Δt in the denominator rather than $\overline{\theta} - \underline{\theta}$, at least for interior $\underline{t}_j \in \left[\underline{\theta}, \overline{\theta} + \Delta t\right]$:

$$\frac{\partial}{\partial \underline{t}_j} \Pr\left\{ j\left(3\right) = j \left| \underline{t}_j \in \left[\underline{\theta}, \overline{\theta} + \Delta t\right], \pi \right. \right\} = \frac{1}{\Delta t}\left(\pi - \frac{1}{2}\right).$$

Respective denominators in the foot-dragging Equations (A-2.4) and (A-2.5) should also now have Δt instead of $\overline{\theta} - \underline{\theta}$ for interior solutions $\underline{t}^* \in \left[\underline{\theta}, \overline{\theta} + \Delta t\right]$. For example, in the baseline scenario of no physician reporting but triage nurse commitment, the equilibrium foot-dragging by low-type physicians is

$$\underline{t}^* = \underline{\theta} + \frac{K_P}{2\Delta t}\left(\pi - \frac{1}{2}\right).$$

This slight modification communicates the intuition that as the expected future work increases (decreases), the marginal temptation to foot-drag increases (decreases) because the certainty of receiving a new patient within each infinitessimal unit of time around discharge is greater (smaller).

A-3 Self-managed System

In this section, I will analyze foot-dragging and patient assignment in the self-managed system. I assume the same physician utilities and information structure (i.e., that they observe each other's θ_j with probability $\psi > 0$) as I did for the nurse-managed system. The only difference will be that the two physicians, not a triage nurse, are responsible for deciding who gets the new patient. In what follows, I will also consider analagous cases in which physicians may or may not report θ_j to each other and in which they may or may not be able to commit to an assignment policy function based on censuses. The key difference between results for all of these cases and corresponding results for the nurse-managed system derives from physicians *both* observing each other's θ_j with probability $\psi > 0$ *and* being able to use that information in patient assignment.

The assignment of patients by the physicians themselves deserves further mention. In the case in which physicians are unable to commit to an assignment policy, I represent assignment as a non-cooperative bargaining game in which physicians can choose to see the new patient. At any time after the new patient has arrived, as long as the patient remains unchosen, either

[7] I maintain the assumption of a single new patient, so that I do not have to consider capacity constraints and physician strategy for subsequent patients, in order to keep the model simple. I also focus on interior solutions so that this single patient is relevant for comparative statics.

physician may choose to see the new patient or wait. If one physician chooses the patient, he gets the patient with probability 1. If they both choose the patient at the same time, they each get the patient with probability $\frac{1}{2}$. This game is very much motivated by Rubinstein's (1982) non-cooperative bargaining game in which two players with complete information about each other's costs bargain over a good that declines in value over time. Aided by a setup that is simpler than Rubinstein's, I will also extend its analysis to consider incomplete information about peer types.

In the other case in which physicians can commit to an assignment policy, I represent an assignment policy function that is quite similar to the one I defined previously for the nurse-managed system. Here the probability of assignment to physician 1 is a function of censuses and potential observations of type $o_j \in \{\emptyset, \theta_j\}$, where recall that both types are observed with probability ψ. That is, the policy function takes the form $\pi_S(c_1, c_2, o_1, o_2)$. It is easy to see that physicians should commit to $\pi_S\left(c_1, c_2, \underline{\theta}, \overline{\theta}\right) = 1$ for all c_1 and c_2. Similar to before, the one relevant parameter to be solved for is $\pi_S \equiv \pi_S(0, 1, \emptyset, \emptyset)$. Commitment in this case involves physicians jointly determining an assignment rule to maximize expected utility prior to knowing their types.

A-3.1 No Physician Reporting, No Policy Commitment

I first consider the case in which physicians can neither report their types nor commit to an assignment policy. In order to allow a war of attrition, I need to elaborate on the cost of waiting to treat a patient who has arrived. While I consider this extended physician utility, which accounts for the cost of delay in assignment, it will be obvious that this utility function is equivalent to the baseline utility function (A-1.1) used in the rest of the conceptual framework (and in the main text) when there is no delay.

A-3.1.1 Extended Physician Utility

Denote τ as the time that has elapsed since the new patient arrived. Assume that both physicians incur a cost of having the new patient remain untreated at each time τ, increasing over τ, and also assume now that the cost of getting the new patient also increases over τ but not as quickly. I then extend physician utility as

$$u_j^P = -\left(t_j - \theta_j\right)^2 - W_j\left(\theta_j; \tau^*, j(3)\right) - K_P\left(\theta_j\right) \mathbb{I}\left\{j(3) = j\right\}. \tag{A-3.1}$$

The new second term $W_j(\cdot)$ is a generalized cost of waiting. Denoting τ^* as the elapsed time that it took for the new patient to be chosen by someone, and $j(3)$ as the physician who chose the patient, W_j is simply integrated over each τ.

$$W_j\left(\theta_j; \tau^*, j(3)\right) = \int_0^{\tau^*} \omega_j\left(\tau; \theta_j, \tau^*, j(3)\right) d\tau,$$

$\omega_j(\tau;\theta_j,\tau^*,j(3))$ is a flow cost that depends on τ and whether physician j received the patient at τ or not:

$$\omega_j(\tau;\theta_j,\tau^*,j(3)) = \begin{cases} w\tau, & \text{if } \tau \neq \tau^* \\ k_0(\theta_j) + k\tau, & \text{if } \tau = \tau^*,\ j = j(3) \\ 0, & \text{if } \tau = \tau^*,\ j \neq j(3). \end{cases}$$

$w\tau$ is the flow cost of waiting imposed on both physicians as long as the patient remains unchosen. $k_0(\theta_j)+k\tau$ is the initial flow cost of seeing the patient, where I denote $\underline{k}_0 = k_0(\underline{\theta})$ and $\overline{k}_0 = k_0(\overline{\theta})$ for brevity and assume $\underline{k}_0 < \overline{k}_0$. Note that since $\tau = \tau^*$ with mass of 0, I can simplify the cost-of-waiting integral to $W_j(\tau^*) = \frac{1}{2}w\tau^{*2}$.

I assume that the flow costs of seeing that patient, starting with the initial cost $k_0(\theta_j)$ if the patient is seen immediately, integrate to $K_P(\theta_j)$. This ensures that the extended utility in (A-3.1) reduces to Equation (A-1.1) when there is no delay in assignment.[8] To ensure single crossing, I impose that k and the continued flow cost of seeing the new patient are both less than w. Finally I assume that physicians are responsible for taking care of patients up to $t = \overline{\theta} + \overline{k}_0/(w-k)$ to ensure that any patient arriving up to $t = \overline{\theta}$ could be chosen by a physician.

A-3.1.2 Non-cooperative Assignment by Physicians

In order to analyze assignment by physician choice, I consider four different cases. The first two cases occur when physicians observe each other's types, which happens with probability ψ. Case 1 is that physicians observe that one is type $\underline{\theta}$ and the other is type $\overline{\theta}$ and proceeds quite similarly to Rubinstein's (1982) setup with complete information. Physicians will infer that a low-type physician will weakly prefer choosing the patient at $\underline{\tau} = \underline{k}_0/(w-k)$, while a high-type physician will weakly prefer choosing the patient at $\overline{\tau} = \overline{k}_0/(w-k)$ and will strictly prefer waiting at $\underline{\tau} < \overline{\tau}$. Therefore, by the subgame where physicians choose to see the patient at $\underline{\tau}$, a low-type physician will be assigned the patient with probability 1. Given that there is a cost to waiting, a low-type physician will prefer to see the patient immediately than to wait until $\underline{\tau}$. So in equilibrium, a low-type physician will choose the patient at $\tau^* = 0$ with probability 1.

Case 2 is that physicians observe that they are both the same type $\theta \in \{\underline{\theta}, \overline{\theta}\}$. There exists no pure strategy Nash equilibrium here. To see this, if there exists a Nash equilibrium, under symmetry both physicians should choose the patient at some $\tau^*(\theta)$. But one physician would always be strictly better off by waiting and letting the other physician choose the patient. There does exist a mixed strategy Nash equilibrium though. I generally denote the probability with which each physician will choose the patient at each elapsed time τ as $q(\theta;\theta,\tau)$, where the first argument is the type θ_j of the index physician (in this case shared) and the second argument

[8] This is automatic in the nurse-managed system and assumed in the case of physician commitment to a policy function in the self-managed system. I could also consider that the triage nurse might delay the assignment of patients to physicians if she can gain more information about their types. However, this would not add any useful intuition for the nurse-managed system, and of course, the setup of physicians starting with one patient each is itself an abstraction.

denotes the observation $o_{-j} \in \{\theta_{-j}, \emptyset\}$ of the peer's type (in this case observed). To satisfy the conditions for indifference between choosing and not choosing,

$$q(\theta; \theta, \tau) = \frac{\max\{0, w\tau - (k_0(\theta) + k\tau)\}}{w\tau - (k_0(\theta) + k\tau)/2}, \tag{A-3.2}$$

From Equation (A-3.2), note that $q(\theta, \tau) > 0$ only when the flow cost of waiting is inefficiently greater than the flow cost of initially treating the patient, i.e., $\tau > k_0(\theta)/(w-k)$. Also, although $q(\theta, \tau)$ increases with τ, $\lim_{\tau \to \infty} q(\theta) = (w-k)/(w-k/2) < 1$.

The next two cases occur in the subgame, occurring with probability $1-\psi$, in which physicians do not observe each other's types. As in the nurse-managed system, a high-type physician will still discharge his patient at $t = \overline{\theta}$. Suppose that a low-type physician discharges his patient at $t = \underline{t}^*$. Types can then be perfectly deduced only during times $t \geq \underline{t}^*$. So prior to $\underline{t}^*$, we have a game of incomplete information, and if the new patient arrives at $t_a < \underline{t}^*$, physicians will have an incentive to wait because types are unknown and therefore certain times for patient choice (i.e., Case 1) are impossible. Thus, during any $t \in T_\emptyset[t_a, \underline{t}^*]$, physicians must decide whether to choose patients while peer types are unknown. Case 3 considers the choice equilibrium strategy for a low-type physician during this time. As in Case 2, he will purely wait if $\tau \leq \underline{k}_0/(w-k)$ and will engage in a mixed strategy if $\tau > \underline{k}_0/(w-k)$. Similar to Equation (A-3.2),

$$q(\underline{\theta}; \emptyset, \tau) = \min\left\{1, \frac{\max\{0, w\tau - (\underline{k}_0 + k\tau)\}}{p(w\tau - (\underline{k}_0 + k\tau)/2)}\right\}. \tag{A-3.3}$$

Equation (A-3.3) shows that a low-type physician is actually more likely to choose the new patient when he does not know that his peer's type relative to when he knows his peer is also low-type (but obviously less likely than when he knows his peer is high-type). Also, if the probability p of having a low-type peer is low enough (i.e., if $p < (w-k)/(w-k/2)$), he may even choose the patient with certainty at some point.

Case 4 considers the strategy for a high-type physician during the same $T_\emptyset = [t_a, \underline{t}^*]$. For a given elapsed time since arrival τ, note that both high-type and low-type physicians cannot be engaging in a mixed strategy. This would require

$$\frac{w\tau - (\underline{k}_0 + k\tau)}{w\tau - (\underline{k}_0 + k\tau)/2} = pq(\underline{\theta}; \emptyset, \tau) + (1-p)\, q(\overline{\theta}; \emptyset, \tau) = \frac{w\tau - (\overline{k}_0 + k\tau)}{w\tau - (\overline{k}_0 + k\tau)/2},$$

which is impossible for finite τ, given $\underline{k}_0 < \overline{k}_0$. So in equilibrium, high-type physicians will wait until after the time when low-type physicians should have chosen the new patient with certainty, which is implied in (A-3.3) to be

$$\underline{\tau}^* = \frac{\left(1 - \frac{1}{2}p\right)\underline{k}_0}{(1-p)\,w - \left(1 - \frac{1}{2}p\right)k}.$$

Note that types will then be revealed at $t_R = \min\left(\underline{t}^*, t_a + \underline{\tau}^*\right)$.

Finally, note that if the patient is still unchosen at t_R when physician types are revealed, then by definition, both physicians must be high type, and they will know this. They should mix with probability as defined in (A-3.2) as in Case 2. They could wait until some time after t_R until they start mixing, or they could immediately start mixing at t_R, depending on whether $t_R - t_a > \overline{k}_0 / \left(w - k\right)$. As in Case 2, each physician will never choose the patient with certainty.

A-3.1.3 No Incentive for Foot-dragging

Considering all 4 cases in the section above, it is clear that low-type physicians can never hope to have the new patient assigned to a high-type peer. This is true even when types are unknown by peers, because as shown in Case 4, high-type physicians will still wait until any low-type physician would have chosen the patient with certainty before choosing the patient with any positive probability.

Another point that is obvious from the above analysis of physician assignment without policy commitment is that physicians will often delay choosing patients, if they know that they are equally busy or if they are unsure how busy their peer is. In fact, under this delay, patients may sometimes never be chosen, given that there is no commitment to choose patients eventually. The delay in choosing patients represents the channel of "free-riding." However, this represents an *ex ante* utility loss at the stage that low-type physicians decide $\underline{t}^*$, since physicians only start placing some positive probability on choosing the new patient when they would have chosen the patient anyway if there were no peer.

Proposition A-6. *In the Perfect Bayesian Equilibrium in the self-managed system with no physician reporting or commitment to an assignment policy, there will be no foot-dragging, i..e,* $\underline{t}^* = \underline{\theta}$. *If the patient is chosen and if there is a low-type physician, assignment will always be to a low-type physician. However, physicians will wait to choose the new patient (free-ride) if they are of the same type.*

Low-type physicians have no incentive to conceal their types by foot-dragging for two reasons: First, low-type physicians can never hope to have a potential high-type peer choose the new patient before them. Second, concealing that they are low-type (foot-dragging) only leads to free-riding, which represents an *ex ante* utility loss. As in the nurse-managed system, assignment is completely *ex post* efficient in the sense that patients are never assigned to physicians with lower censuses when censuses differ. In addition, given that there is no foot-dragging, this also implies *ex ante* efficiency. However, there is a new "assignment" inefficiency of free-riding in that physicians may delay seeing patients and sometimes not even get to see them despite preferring to had there been no peer. In this model, free-riding only occurs when physicians are of the same type, since types can always be inferred by the time the new patient arrives at $t \in \left[\underline{\theta}, \overline{\theta}\right]$, as low-type physicians never foot-drag.[9]

[9] In practice, types may not be perfectly deduced (which would also happen in the model if the new patient

A-3.1.4 Remark on Continuous Types

I will conclude this subsection with a remark on continuous types in order to show that with continuous types the intuition for the stark result in Proposition A-6 does not hold. To see this, first note that with continuous types, even if types are unknown, the probability that a physician has the same type as his peer has mass 0. Thus, physicians should choose patients with pure strategies in equilibrium.

When physician do not observe each other's types (with probability $1-\psi$), physicians then use peer signals to infer θ_{-j}, and by subgame perfection physician j will choose the new patient at $\tau=0$ if and only if $c_j > c_{-j}$. We now have a similar situation as in the traditional system with no triage nurse commitment, in Section A-2.3, where there is a first-order gain to foot-dragging. So in equilibrium, there should be no free-riding but positive foot-dragging. However, recall that with probability ψ, physicians observe each other's types. In this case, the physician j will choose the new patient at $\tau=0$ if and only if $\theta_j > \theta_{-j}$, regardless of c_j and c_{-j}. Because of this, physicians will foot-drag less than they would have under the traditional system with no triage nurse commitment.

A-3.2 No Physician Reporting, Policy Commitment

From the analysis in Section A-3.1, it is clear that without physician commitment to an assignment policy, there could be large welfare losses in the form of free-riding and patients going untreated. This suggests scope for improvement by committing to an assignment policy. As a practical rationale, physicians often divide work before they can deduce each other's true workloads if new patients need to be seen in a timely manner.

Introduced above, the policy function takes the form $\pi_S\left(c_1,c_2,o_1,o_2\right)$, where o_j is type of physician j if observed and null otherwise. The following are obvious: $\pi_S\left(c_1,c_2,\underline{\theta},\overline{\theta}\right)=1$ for all c_1 and c_2; $\pi_S\left(c_1,c_2,\theta_1,\theta_2\right)=\frac{1}{2}$ for $\theta_1=\theta_2$ and all c_1 and c_2; and $\pi_S\left(c_1,c_2,\emptyset,\emptyset\right)=\frac{1}{2}$ for $c_1=c_2$.[10] As in the nurse-managed system, the assignment policy can then be represented by a single parameter $\pi_S\equiv\pi_S\left(0,1,\emptyset,\emptyset\right)$. In equilibrium, physicians choose the optimal assignment policy π_S^* at time $t=0$, given physician discharge strategies $\underline{t}^*$ and $\overline{t}^*$ for low- and high-type physicians, respectively. Given this assignment policy, physicians choose the optimal discharge strategies $\underline{t}^*$ and $\overline{t}^*$.

Proposition A-7. *In the Perfect Bayesian Equilibrium in the self-managed system with no physician reporting but with commitment to an assignment policy, if as $\psi>0$, $\Delta K_P>\overline{K}_N$, and $D=1$, then there will be less foot-dragging and more* ex post *efficient assignment than in the nurse-managed system with no physician reporting but triage nurse commitment.*

could arrive at $t<\underline{\theta}$), and this could support free-riding. However, as shown in the main paper, free-riding does not appear to be significant empirically, which suggests that physicians can commit to an assignment policy or have sufficient information about each other's types (either by censuses or observations of true workload).

[10] For simplicity and for consistency with the nurse-managed system, I assume that patients are immediately assigned under this policy. See also footnote (8).

The scenario for no physician reporting and policy commitment is analyzed similarly as in the corresponding scenario in the nurse-managed system. For any given policy π_S, a low-type physician will discharge his patient at time

$$\underline{t}^* = \underline{\theta} + (1-\psi)\frac{\underline{K}_P}{2\left(\overline{\theta}-\underline{\theta}\right)}\left(\pi_S - \frac{1}{2}\right). \tag{A-3.4}$$

Note again the similarity between Equations (A-2.5) and (A-3.4). The only difference is that the second term is multiplied by $1-\psi$, because with probability ψ, foot-dragging will have no effect on assignment. Self-management – both the observation of true workload and the use of this information in assignment – decreases foot-dragging relative to the nurse-managed system.

It now follows that the physicians will choose a policy function that takes Equation (A-3.4) into consideration in order to maximize their *ex ante* utilities, before they have received their initial patients. They expected to be type $\underline{\theta}$ with probability p and type $\overline{\theta}$ with probability $1-p$, and they maximize

$$\begin{aligned}\mathbb{E}\left[u^P\left(\pi_S\right)\right] \quad = \quad & -p\left(\underline{t}^* - \underline{\theta}\right)^2 - (1-p)^2 \Delta K_P - \\ & 2p\left(1-p\right)\left(1-\psi\right)\Delta K_P\left[\frac{1}{2}\left(\frac{\underline{t}^*-\underline{\theta}}{\overline{\theta}-\underline{\theta}}\right) + \left(1-\pi_S\right)\left(\frac{\overline{\theta}-\underline{t}^*}{\overline{\theta}-\underline{\theta}}\right)\right],\end{aligned} \tag{A-3.5}$$

where I conveniently transform the expected utility by subtracting $\underline{K}_P$ and using $\Delta K_P \equiv \overline{K}_P - \underline{K}_P$. Again, note the similarity with Equation (A-2.6), with two differences. First, with probability ψ, the policy function is irrelevant for assignment, so $1-\psi$ appears in the third term.[11] Second, instead of $\overline{K}_N$, the analogous parameter ΔK_P from the physician's utility function is used because physicians are the ones making patient assignment.

Ex ante, physicians would like to avoid assigning the new patient to a busier physician because it is more costly, by ΔK_P, for that physician to deal with that patient. However, *ex post*, once physicians know their types, low-type physicians have the moral hazard to avoid the new patient. Commitment to a policy function allows physicians *ex ante* to balance their desire for proper assignment with the knowledge that proper assignment will cause costly moral hazard.

Maximizing (A-3.5) with respect to π_S, after substituting (A-3.4) for $\underline{t}^*$, yields the optimal policy function

$$\pi_S^* = \frac{1}{2} + \frac{4p\left(1-p\right)\Delta K_P\left(\overline{\theta}-\underline{\theta}\right)^2}{\left(1-\psi\right)\underline{K}_P\left(p\underline{K}_P + 4p\left(1-p\right)\Delta K_P\right)}. \tag{A-3.6}$$

The differences between (A-2.7) and (A-3.6) are twofold. First, the denominator is multiplied by $1-\psi$, which reflects the fact that foot-dragging is lessened by the possible observation of true workload, and which improves the efficiency of assignment even when true workload is

[11] The third term also represents the efficiency loss with respect to misassignment under any policy commitment. It is larger with small ψ, p close to 1/2, and large ΔK_P. On the other hand, the efficiency loss with no policy commitment is larger with large k, large w, or p close to 0 or 1. Depending on these parameters, physicians may opt for no policy function even if they can commit to one.

not observed. Second, ΔK_P replaces $\overline{K}_N$. Thus, as long as $\psi > 0$ and $\Delta K_P > \overline{K}_N$, then $\pi^*_{N|D=1} < \pi^*_S$.

Although I cannot definitively say that one is bigger than the other, it is likely that $\Delta K_P > \overline{K}_N$. To see this, notice that the triage nurse's utility function in Equation (A-2.1) can include the outcomes of *both* physicians. $\overline{K}_N$ represents the amount that she values assignment of the third patient compared to the amount that she values the average of outcomes for the first two patients. On the other hand, in the physician utility function in Equation (A-1.1), the cost of receiving another patient is scaled relative to the outcome of a single patient. So if the triage nurse had similar preferences as both physicians, we should have $\Delta K_P \approx 2\overline{K}_N$. For both of these reasons, π^*_S again should be greater than π^*_N, meaning that the self-managed system improves the efficiency of assignment relative to the nurse-managed system.

A-3.3 Physician Reporting, Policy Commitment

Finally, in the case of physician reporting and policy commitment, it suffices to consider how the observation of true workloads between peers and its use in the policy function modifies incentive compatibility constraints for a low-type physician.

The standard incentive compatibility constraint, assuming "full" lying, is

$$-\left(t\left(\overline{\theta}\right)-\underline{\theta}\right)^2-\frac{1}{2}\left(1-\psi\right)\underline{K}_P \leq -\pi_S\left(1-\psi\right)\underline{K}_P. \tag{A-3.7}$$

Even if he mimics a high-type physician, he will still be observed as a low-type physician with probability ψ, in which case mimicry was useless. This relaxes the incentive compatibility constraint that was originally (A-2.9) and allows a higher π_S to support truth-telling.

However, as in the nurse-managed system, the intertemporal incentive compatibility constraint will be binding. Recall that this constraint, shown in Equation (A-2.11) for the nurse-managed system, derives from the concern that low-type physicians can lie about their type and then discharge their patient earlier than $t\left(\overline{\theta}\right)=\overline{\theta}$. The constraint in the self-managed system is

$$-\left(\hat{t}-\underline{\theta}\right)^2-\left(1-\psi\right)\underline{K}_P\left[\frac{1}{2}\left(\frac{\hat{t}-\underline{\theta}}{\overline{\theta}-\underline{\theta}}\right)+\frac{\overline{\theta}-\hat{t}}{\overline{\theta}-\underline{\theta}}\right] \leq -\pi_S\left(1-\psi\right)\underline{K}_P, \tag{A-3.8}$$

where $\hat{t}$ is the time that a low-type physician reveals that he was lying when he initially reported that he was high-type. Again, with probability ψ, this lie will not pay off, which relaxes the incentive compatibility constraint to allow a higher π_S. Thus, assignment will be more efficient in the self-managed system compared to the nurse-managed system in this case as well.

Proposition A-8. *In the Perfect Bayesian Equilibrium in the self-managed system with physician reporting and commitment to an assignment policy, assignment will be more* ex ante *(and* ex post*) than in the nurse-managed system with physician reporting and triage nurse commitment. There is still no foot-dragging, given truth-telling and discrete types.*

A-4 Continuous Types

In this section, I relax the baseline two-type model to consider a continuum of types in the nurse-managed system with physician reporting. The purpose of this extended model is to communicate the intuition that, in contrast to the two-type analysis in Section A-2.4 with physician reporting, there will still be foot-dragging and inefficient assignment, as long as types are sufficiently rich. As discussed in Section A-2.4, this is much more reflective of reality, particular in settings in which information is complex, the very source of physician discretion in the first place. While I restrict attention to the nurse-managed system, similar intuition follows for the self-managed system.

The model is identical to the model outlined in Section A-1 except for two differences: First, each of the two physicians can be of type $\theta_j \in \left[\underline{\theta}, \overline{\theta}\right]$, drawn from some distribution which I do not specify.[12] Second, because types lie in a continuum, I allow for a more flexible triage nurse policy function that takes the form of $\pi\left(\theta_1, \theta_2, t\right)$, which is the flow probability of the physician 1 receiving a patient who arrives at time t when the types of physicians 1 and 2 are θ_1 and θ_2, respectively. Of course, this policy function cannot be reduced to a single parameter, because both θ_1 and θ_2 are continuous. In addition, I allow for the fact that the optimal policy may have a non-constant flow-rate for a given θ_1 and θ_2. The sufficient statistic from the standpoint of physician and triage nurse utility is

$$P\left(\theta_1, \theta_2\right) \equiv \frac{1}{\overline{\theta} - \underline{\theta}} \int_{\underline{\theta}}^{\overline{\theta}} \pi\left(\theta_1, \theta_2, t\right) dt,$$

which is the cumulative probability over time that physician 1 will receive the new patient, given that patient arrival is uniformly distributed and that assignment is immediate upon arrival. The reason I allow for the flexible time-dependent flow is to specifically consider "intertemporal feasibility" constraints in which $\pi\left(\theta_1, \theta_2, t\right) \leq 1$, for all θ_1, θ_2, t, which I have noted in Section A-2.4 as feature different from the standard screening problem.

I will analyze this model as follows: First, I will formalize the intertemporal feasibility constraints as incentive compatibility constraints for truth-telling. That is, physicians should have no incentive to discharge their patient at time $t < t\left(\hat{\theta}_j\right)$, where $\hat{\theta}_j$ is their reported type, which may be different than their true type θ_j. Second, as in the standard mechanism design problem, I will show that I can restrict attention to local incentive compatibility constraints in which physicians have no incentive to report the type continuously adjacent to theirs, conditional on the previous requirement that they follow the discharge time required by that report. Third, I will show by perturbation arguments that the optimal triage nurse policy function is continuous and strictly increasing. This implies that there will be positive foot-dragging and *ex post* inefficient

[12] The distribution is not important for this analysis, which shows positive foot-dragging and inefficient assignment, but it would be necessary to consider for an analysis that computes the optimal assignment function in closed form.

assignment in the sense that $P(\theta_j, \theta_{-j}) < 1$ for any $\theta_j < \theta_{-j}$. That is, in the remainder of this section, I will show the following:

Proposition A-9. *In the Perfect Bayesian Equilibrium in the nurse-managed system with physician reporting and triage nurse commitment, for a continuum of physician types distributed along* $\left[\underline{\theta}, \overline{\theta}\right]$*, there will be positive foot-dragging such that* $t(\theta) > \theta$ *for all* $\theta < \overline{\theta}$ *and* ex post *inefficient assignment such that* $P(\theta_j, \theta_{-j}) < 1$ *for any* $\theta_j < \theta_{-j}$*.*

A-4.1 Intertemporal Feasibility

As in any truth-telling equilibrium, physicians must not have the incentive to misreport their types. This setting in particular requires me to address the possibility of a physician reporting $\hat{\theta}_j$, in order to get some flow probability $\pi\left(\hat{\theta}_j, \theta_{-j}, t\right)$ of assignment for the new patient, but discharging the current patient at some time $t_j < t\left(\hat{\theta}_j\right)$. If physicians can receive a lower $P\left(\hat{\theta}_j, \theta_{-j}\right)$ by reporting $\hat{\theta}_j > \theta_j$ but not have to keep their patient as long as would be required by $t\left(\hat{\theta}_j\right)$, then they could be strictly better off. The reason for this departure from the standard screening model is that the cumulative probability $P(\theta_1, \theta_2)$ is not given by the triage nurse in a lump sum but rather over time. Thus, there is an "intertemporal feasibility" constraint in that the triage nurse may not be able to take back (in terms of lower probability of assignment) what she has already given in the past. More precisely, this constraint derives from the fact that even punishment policy functions are limited by $\pi(\theta_1, \theta_2, t) \leq 1$, for all θ_1, θ_2, and t.

In order to account for this, I simply consider that, for the standard mechanism design problem to work here, I need physicians to have no incentive to report $\hat{\theta}_j$ and then discharge their patient at time $t_j < t\left(\hat{\theta}_j\right)$ as opposed to reporting $\tilde{\theta}_j$ and discharging their patient at the same time $t_j = t\left(\tilde{\theta}_j\right)$. That is, if the physician is planning to discharge his patient at some time t_j, he may as well report the type that corresponds to that time. Note that this incentive compatibility does not yet require that the physician prefers to report $\tilde{\theta}_j = \theta_j$, which I consider later. In order to sustain this aspect of truth-telling, I assume that if the triage nurse catches a physician lying *and* deviating by $t_j < t\left(\hat{\theta}_j\right)$, then she will punish him by assigning him the new patient with probability 1 – but no more by the feasibility constraint – if the new patient arrives at $t \in \left[t_j, \overline{\theta}\right]$.

The incentive compatibility constraint implied by intertemporal feasibility then can be stated as

$$\begin{aligned} P\left(\tilde{\theta}_j, \theta_{-j}\right) &\equiv \frac{1}{\overline{\theta} - \underline{\theta}} \int_{\underline{\theta}}^{\overline{\theta}} \pi\left(\tilde{\theta}_j, \theta_{-j}, t\right) dt \\ &\leq \frac{1}{\overline{\theta} - \underline{\theta}} \left[\int_{\underline{\theta}}^{t(\tilde{\theta}_j)} \pi\left(\hat{\theta}_j, \theta_{-j}, t\right) dt + \overline{\theta} - t\left(\tilde{\theta}_j\right)\right], \ \forall \hat{\theta}_j < \tilde{\theta}_j, \theta_{-j} \end{aligned}$$

where I ignore the potential cost of discharging a patient at time $t\left(\tilde{\theta}_j\right)$, since this is held constant. This incentive compatibility constraint can equivalently be stated as

$$\int_{t(\theta_j)}^{\underline{\theta}} 1-\pi\left(\theta_j,\theta_{-j},t\right)dt \geq \int_{\underline{\theta}}^{t(\theta_j)} \pi\left(\theta_j,\theta_{-j},t\right)-\pi\left(\hat{\theta}_j,\theta_{-j},t\right)dt,\ \forall\hat{\theta}_j<\theta_j,\theta_{-j}. \tag{A-4.1}$$

This formalizes the intuition that the triage nurse requires scope for punishment in the policy function in order to prevent physicians from misreporting their types.

A-4.2 Local Incentive Compatibility

The remainder of the analysis proceeds similarly to the standard screening mechanism design. The triage nurse offers a menu of choices $\{(t(\theta_1,\theta_2),P(\theta_1,\theta_2))\}$ to physicians. Physicians simultaneously report θ_1 and θ_2, and by the Revelation Principle, they report truthfully.[13] The triage nurse's problem is to choose a menu that maximizes her expected utility subject to physician truth-telling.

I first show that the single-crossing condition holds. Physician utility remains the same as in the baseline model, but I account for continuous types by allowing the cost incurred by being assigned the new patient, $K_P(\theta)$, to be a continuous and differentiable function such that $K_P'(\theta)\geq 0$. It is easy to show that

$$\frac{\partial}{\partial\theta_j}\left[-\frac{\partial u_j^P/\partial P\left(\theta_j,\theta_{-j}\right)}{\partial u_j^P/\partial t_j}\right]>0$$

as long as $t>\theta$ and $K_P'(\theta)\geq 0$. The former condition that $t>\theta$, equivalent to foot-dragging, will be shown later to hold in equilibrium. I will assume the latter more strictly by $K_P'(\theta)>0$.

Given the single-crossing condition, the set of incentive compatibility constraints,

$$-\left(t\left(\theta_j,\theta_{-j}\right)-\theta_j\right)^2-K_P\left(\theta_j\right)P\left(\theta_j,\theta_{-j}\right)\geq-\left(t\left(\hat{\theta}_j,\theta_{-j}\right)-\theta_j\right)^2-K_P\left(\theta_j\right)P\left(\hat{\theta}_j,\theta_{-j}\right),$$

for all $\theta_j,\theta_{-j},\hat{\theta}_j$, can be summarized as a monotonicity condition and local incentive compatibility constraints. The monotonicity condition is

$$\frac{\partial P\left(\theta_j,\theta_{-j}\right)}{\partial\theta_j}\leq 0, \tag{A-4.2}$$

and the local incentive compatibility constraints are

$$-2\left(t\left(\theta_j,\theta_{-j}\right)-\theta_j\right)\frac{\partial t\left(\theta_j,\theta_{-j}\right)}{\partial\theta_j}=K_P'\left(\theta_j\right)\frac{\partial P\left(\theta_j,\theta_{-j}\right)}{\partial\theta_j}, \tag{A-4.3}$$

[13] In the following analysis, I assume that physicians know each other's types, but the intuition follows if they only act according to expected values of their peer's type.

simply stated by setting the derivative of the physician utility function equal to 0.[14]

A-4.3 Optimization Problem

I will now analyze the triage nurse's optimization problem subject to (A-4.1) and (A-4.3). Rather than solve her problem in closed form, I can obtain my results – that there will be foot-dragging and *ex post* inefficient assignment in equilibrium – by simple perturbation arguments.

I first maintain the assumption that the triage nurse's utility function takes the form as stated in (A-2.1), and I similarly assume that $K_N(\theta)$ is continuous and twice differentiable and that furthermore $K_N^{'}(\theta) > 0$. It then follows that she will offer menu options $t(\theta_j, \theta_{-j})$ and $P(\theta_j, \theta_{-j})$ that are continuous and differentiable in θ_j for all θ_j and θ_{-j}. To see this, suppose that her optimal assignment policy function $P(\theta_j, \theta_{-j})$ is discontinuous at $\theta_j = \theta_1$ and some $\theta_{-j} = \theta_2$, such that $\lim_{\varepsilon \to 0} P(\theta_1 - \varepsilon, \theta_2) - P(\theta_1 + \varepsilon, \theta_2) = \Delta > 0$. If this is the case, then no physician whose type $\theta_j \in (\theta_1 - \epsilon, \theta_1)$, for some $\epsilon > 0$, will truthfully reveal his type. The triage nurse cannot maintain truth-telling over some interval of types of strictly positive measure and therefore cannot implement her policy function, which is a contradiction.

Next, note that it is never optimal to have $t(\theta) < \theta$, as it reduces the discharge time away from what is socially optimal and only reduces the utility of the physician when truth-telling, which can only make the truth-telling constraint more costly. Given this, I can show that the optimal assignment policy $P(\theta_j, \theta_{-j})$ is strictly decreasing in θ_j. To see this, suppose that that the optimal policy is such that $P(\theta_j + \varepsilon, \theta_{-j}) \geq P(\theta_j, \theta_{-j})$ for some small $\varepsilon > 0$ and for some θ_j, θ_{-j}. However, the triage nurse can strictly increase her utility and maintain truth-telling by increasing $P(\theta_j, \theta_{-j})$ by some small $\epsilon > 0$ while keeping her set of $\{t(\theta_j, \theta_{-j})\}$ unchanged if $\partial t(\theta_j, \theta_{-j}) / \partial \theta_j > 0$, or by increasing $P(\theta_j, \theta_{-j})$ by some small $\epsilon > 0$ and concurrently decreasing $t(\theta_j, \theta_{-j})$ by some small $\delta > 0$ if $\partial t(\theta_j, \theta_{-j}) / \partial \theta_j \leq 0$. Note that this satisfies the monotonicity condition in (A-4.2).

Given that $P(\theta_j, \theta_{-j})$ is strictly decreasing in θ_j (by contradiction), $t(\theta) \geq \theta$ (by contradiction), and $K_P^{'}(\theta_j) > 0$ (by assumption), then (A-4.3) implies that $\partial t(\theta_j, \theta_{-j}) / \partial \theta_j > 0$ and $t(\theta_j) > \theta_j$, except for type $\theta_j = \overline{\theta}$, which need not be bound by an incentive compatibility constraint. That is, discharge times increase with the physician's type, regardless of his peer's type, and there is positive foot-dragging in equilibrium for all types $\theta_j < \overline{\theta}$.

Finally, *ex post* inefficient assignment derives from two independent facts shown above, either of which is sufficient. First, the continuity of $P(\theta_j, \theta_{-j})$ guarantees *ex post* inefficient assignment because it is impossible to have

$$P(\theta_j, \theta_{-j}) = \begin{cases} 1, & \theta_j < \theta_{-j} \\ 0, & \theta_j > \theta_{-j} \end{cases}, \tag{A-4.4}$$

[14] An additional condition for the local incentive compatibility constraint to be valid is that the menu options $t(\theta_j, \theta_{-j})$ and $P(\theta_j, \theta_{-j})$ are continuous and differentiable in θ_j for all θ_j and θ_{-j}. This will also be shown below.

for all θ_j, θ_{-j}, without a discontinuity at $\theta_j = \theta_{-j}$. Second, inefficient assignment can also be proven by using intertemporal feasibility alone, as stated in (A-4.1). This constraint implies that it is impossible to have both $P(\theta_j, \theta_{-j}) > P\left(\hat{\theta}_j, \theta_{-j}\right)$ and $P(\theta_j, \theta_{-j}) = 1$, for any $\hat{\theta}_j$, θ_j, and θ_{-j}, and therefore also rules out (A-4.4).

References

Apker, Julie, Larry A. Mallak, and Scott C. Gibson, "Communicating in the Gray Zone: Perceptions about Emergency Physician-hospitalist Handoffs and Patient Safety," *Academic Emergency Medicine*, October 2007, *14* (10), 884–894.

Milgrom, Paul and John Roberts, "An Economic Approach to Influence Activities in Organizations," *The American Journal of Sociology*, 1988, *94* (Supplement), 154–179.

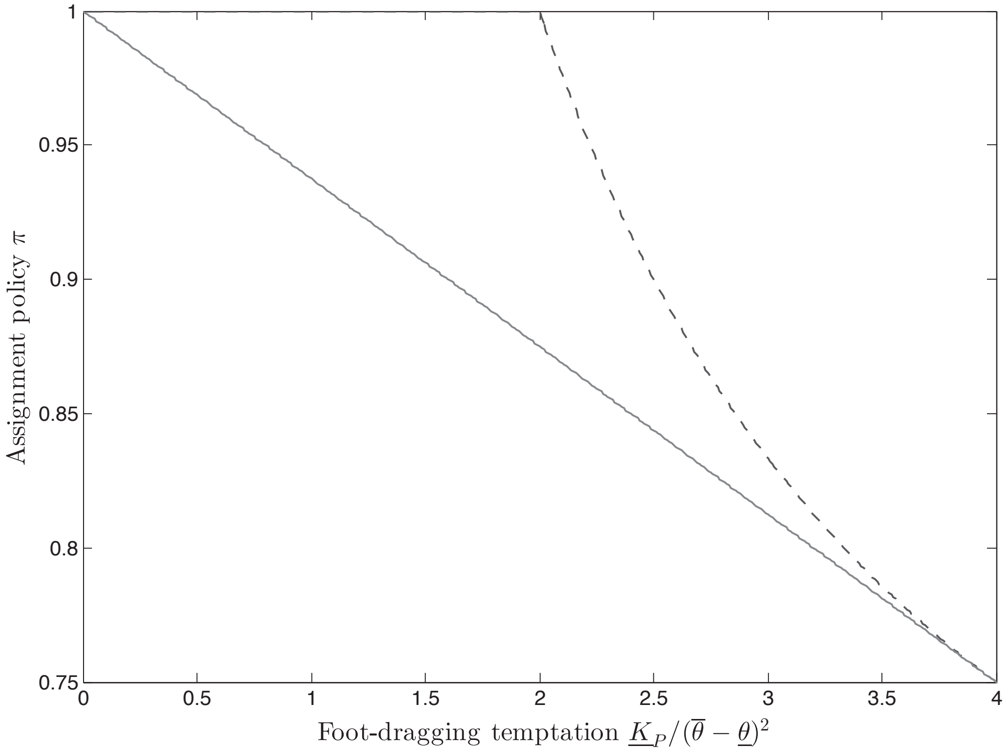
Moore, J. and R. Repullo, "Subgame perfect implementation," *Econometrica*, 1988, *56* (5), 1191–1220.

Myerson, Roger B and Mark A Satterthwaite, "Efficient mechanisms for bilateral trading," *Journal of Economic Theory*, April 1983, *29* (2), 265–281.

Rochet, Jean-Charles and Lars Stole, "The Economics of Multidimensional Screening," in M. Dewatripont, L. Hansen, and S. Turnovsky, eds., *Advances in Economics and Econometrics*, Vol. 35, Cambridge University Press, January 2003, pp. 150–197.

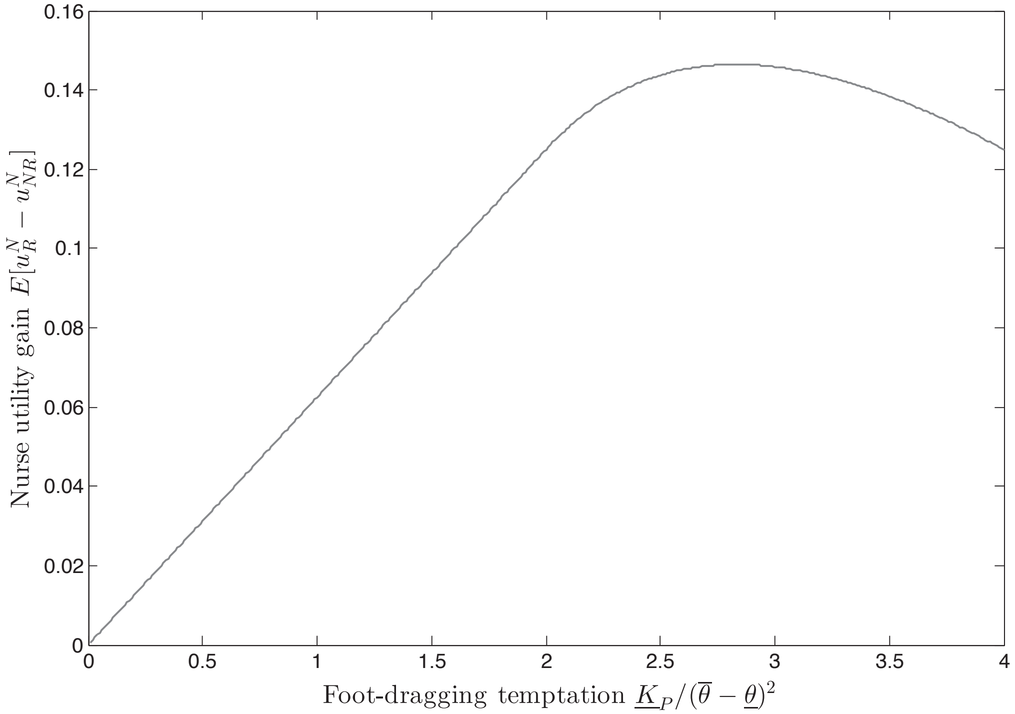
Rubinstein, Ariel, "Perfect Equilibrium in a Bargaining Model," *Econometrica*, 1982, *50* (1), 97–109.

Figure A-2.1: Assignment Policy π Depending on Physician Reporting



Note: This figure shows the assignment policy π, depending on whether physicians can report their types to the triage nurse. Under both cases, I assume that the triage nurse can commit to an assignment policy function. Under no physician reporting, I also consider the triage nurse utility function in which she only cares about patient assignment. On the horizontal axis is a summary statistic for the foot-dragging temptation, $\underline{K}_P/\left(\overline{\theta}-\underline{\theta}\right)^2 \in [0,4]$. Note that when $\underline{K}_P/\left(\overline{\theta}-\underline{\theta}\right)^2 = 4$, a low-type physician should fully foot-drag at $\underline{t}^* = \overline{\theta}$. The dashed line plots the assignment policy when physicians cannot report, $\pi^*_{N|D=0}$ given in (A-2.8). The solid line plots the assignment policy when physicians can report, $\pi^*_R = \pi^*_{IT}$, which equals the assignment policy implied by the intertemporal constraint given in Equation (A-2.12). Note that $\pi^*_{N|D=0} = \pi^*_{IC}$, the assignment policy implied by the standard incentive compatibility constraint with "full" lying, given in Equation (A-2.10); however, the former is a function of censuses, while the latter is a function of reported types.

Figure A-2.2: Efficiency Gain in *ex ante* Assignment with Physician Reporting



Note: This figure shows the gain in *ex ante* assignment efficiency, or the gain in triage nurse utility when she only cares about assignment, that occurs with physician reporting. I compare expected utility for the triage nurse, $\mathbb{E}\left[u^N\right]$, under no physician reporting and under physician reporting. In both cases, her expected utility is simply $\Pr\left\{\theta_{j(3)} = \overline{\theta}\right\}$, normalizing $\overline{K}^N = 1$. Under both cases, I assume that the triage nurse can commit to an assignment policy function. On the horizontal axis is a summary statistic for the foot-dragging temptation, $\underline{K}_P/\left(\overline{\theta}-\underline{\theta}\right)^2 \in [0,4]$. Note that when $\underline{K}_P/\left(\overline{\theta}-\underline{\theta}\right)^2 = 4$, a low-type physician should fully foot-drag at $\underline{t}^* = \overline{\theta}$. On the vertical axis, I plot the difference in *ex ante* assignment efficiency between physician reporting and no reporting.