The Efficiency of Slacking Off: Evidence from the Emergency Department*

David C. Chan†

June 15, 2015

Abstract

Work schedules play an important role in utilizing labor in organizations. In this study of emergency department physicians in shift work, schedules induce two distortions: First, physicians "slack off" by accepting fewer patients near end of shift (EOS). Second, physicians distort patient care, incurring higher costs as they spend less time on patients accepted near EOS. Examining how these effects change with shift overlap reveals a tradeoff between the two. Within an hour after the normal time of work completion, physicians are willing to spend hospital resources eight times more than their market wage to preserve their leisure. Accounting for overall costs, I find that physicians slack off at approximately second-best optimal levels.

JEL Codes: D20, I10, L23, L84, M11, M54

*I am thankful to Amitabh Chandra, Daniel Chen, David Cutler, Mike Dickstein, Joe Doyle, Liran Einav, Bob Gibbons, Jeremy Goldhaber-Fiebert, Jon Gruber, Rob Huckman, Dan Kessler, Eddie Lazear, Sara Machado, Grant Miller, Paul Oyer, Maria Polyakova, Orie Shelef, Jonathan Skinner, and Chris Walters, and seminar participants at AEA/ASSA, ASHEcon, Cornell, Columbia, ETH Zurich, iHEA, MIT, NBER Productivity/Innovation/Entrepreneurship, NBER Summer Institute, RAND, Queen's University, UC Irvine, USC, and University of Toronto for helpful comments and suggestions. I acknowledge early support from the NBER Health and Aging Fellowship, under the National Institute of Aging Grant Number T32-AG000186; the Charles A. King Trust Postdoctoral Fellowship, the Medical Foundation; and the Agency for Healthcare Research and Quality Ruth L. Kirschstein Individual Postdoctoral Fellowship F32-HS021044.

†Address: 117 Encina Commons, Room 217; Stanford, CA 94306. Phone: 650-725-9582. Fax: 650-723-1919. Email: david.c.chan@stanford.edu.

1 Introduction

Canonical models of production consider labor as an input but are silent on two important questions about how firms use workers' time: First, how should worker availability be scheduled? Second, how should work be distributed across workers conditional on availability? This paper analyzes scheduling, a widespread form of coordination in organizations, as a principal-agent problem.[1] By defining boundaries of worker availability, schedules may open a margin for distortionary behavior if the appropriate time to complete work tasks is private information.

In this paper, I theoretically and empirically consider the implications of this problem on work assignment. If workers overvalue their leisure time relative to other consequences of their workplace actions, schedules induce two distortions near end of shift (EOS): First, on an extensive margin, workers "slack off" by accepting fewer tasks than socially optimal. Second, on an intensive margin, workers may rush to complete their work, spending less time than socially optimal on tasks they do accept near EOS. Since workers usually have much more discretion on the intensive margin, the second distortion could be significantly more costly and implies that some slacking off is second-best optimal. While the empirical setting of this paper is in health care, the setting shares characteristics with other time-sensitive and information-rich workplaces:[2] Tasks are uncertain and largely non-contractible; compensation contracts are based on availability (or *minimum quantity* of hours worked); delaying assignment is costly; and worker-task specificity imposes some cost of transferring tasks to other workers once assigned.

Emergency department (ED) shiftwork is well-suited to allow me to estimate the effect of schedules on behavior. Shifts ending at different times allow me to separate effects related to shift work from differences due to the time of day. Shifts of different lengths allow separating

[1] A large and active literature in operations management has viewed scheduling workers as mechanical inputs (e.g., Perdikaki et al., 2012; He et al., 2012; Green, 2004, 1984), including recent investigations that describe how worker throughput responds to environmental features such as "system load" (Kc and Terwiesch, 2009). In economics, a team-theoretic literature (e.g., Marschak and Radner, 1972; Radner, 1993; Garicano, 2000) has taken a similar approach. In practice, algorithmic approaches, e.g., using computerized staffing tools, are widely used by firms (Maher, 2007).

[2] Examples of such workplaces could include large-scale construction, management consulting, and software engineering. A number of online worker scheduling services have emerged, and case studies of client firms across industries can be found, for example, at `https://www.shiftplanning.com/casestudies`. The industry need not be 24-7, although the size of the economy involved in 24-7 activities has grown (Presser, 2003), and the number of workers with non-standard work times has grown to more than two-fifths (Beers, 2000). The key factors involve worker discretion and an assignment decision across workers.

these effects from "fatigue," which I consider to depend on the time since the beginning of shift. Physicians work in virtually all types of shifts. I show that physicians accept fewer patients near EOS. For patients they do accept, I also show that physicians shorten the duration of care ("length of stay") in the ED and increase formal utilization, inpatient admissions, and hospital costs as the time of arrival approaches EOS. I find evidence that differential selection of patient types (i.e., selecting healthier patients near EOS) is negligible compared to the size of the effect on length of stay and in the opposite direction of utilization, admissions, and costs.

To interpret changes in patient care as distortions, I use another source of variation from shift structure: the overlapping time between when a peer arrives on a new shift and when the index physician reaches EOS. My identifying assumption is that, conditional on the volume of work, the time from the beginning of the shift, and the time from the peer's arrival, the EOS should have no bearing on efficiency, since it is merely when physicians *may* go home if work is complete. I show that distortions on the intensive margin of patient care are greatest when physicians have the least time to offload work onto a peer before EOS. In fact, there is no increased utilization or admissions when overlap is four or more hours.

This evidence suggests a policy tradeoff between the extensive and intensive margins of distortion. On the extensive margin, workers "slack off" by accepting fewer patients. While slacking off represents a waste of physicians' time, it reduces workload when time becomes more costly and, on the intensive margin, mitigates physicians inefficiently substituting other inputs for time.[3] Using a structural model based on the connection between workload-adjusted length of stay and hospital costs, I consider a wider range counterfactual policies of patient assignment near EOS, and I find that observed assignment patterns approximately minimize overall costs of physician time, patient time, and hospital resources. Assigning more patients near EOS such that physicians stay an additional hour induces an additional $5,500 in hospital spending per shift; physicians also reveal that they are willing to spend more than $990 in hospital dollars per each hour of leisure saved, which is eight times greater than the market wage.

This paper contributes to two strands of literature. First, a central economic question is how

[3]The idea of time per effective work is related to work by Coviello et al. (2014), who discuss of the effect of dividing time among tasks, although with a single worker who works indefinitely. The time for completing a project mechanically is lower when fewer projects are active because time is divided among fewer projects.

to induce workers to work efficiently, analyzed through the lens of incomplete contracts and the principal-agent problem (Simon, 1947; Hart and Holmstrom, 1987). Following seminal papers that evaluate a manager's second-best optimal policy under hidden action or information, (e.g., Shapiro and Stiglitz, 1984; Aghion and Tirole, 1997; Milgrom and Roberts, 1988), I apply this framework to the design of scheduling and assignment, and I find that work assignment should be lower than first-best near EOS. This paper also contributes to an empirical literature on the relationship between workplace design and productivity.[4] In particular, recent literature suggests that workplaces that grant greater flexibility to workers in how, when, and where work is performed have greater productivity (Ichniowski et al., 1996; Bloom et al., 2014).

Second, this paper sheds empirical light on the balance between extrinsic and intrinsic motivation (e.g., Benabou and Tirole, 2003). While workers no doubt care about their income and leisure, a now-substantial literature in economics recognizes that workers care about the "mission" of their job.[5] In medical care, where information is continuous, multidimensional, and difficult to communicate, it would be extremely difficult to design incentives to provide the right care for patients if physicians only cared about income and leisure. By construction, salaries and schedules provide an environment in which extrinsic motives are muted relative to intrinsic ones, but the boundaries of schedules present a unique opportunity to study the tradeoff between private and intrinsic mission-oriented goals. In this paper, I find that the reduced-form tradeoff depends on the time during shift and quickly grows large in favor of private goals.

The issues I study in this paper are particularly relevant to health care delivery, which has experienced broad changes in the use of labor over the last few decades. Technological advances have caused a proliferation in the diagnostic and therapeutic decisions that should be made in rapid order from a patient's presentation.[6] Further, changes in work and society, including the emergence of dual-earner families, have driven worker preferences for more predictable yet

[4]Relatedly, an interesting set of papers has studied timing distortions in nonlinear contracts such as sales incentive plans and government budgets (e.g., Oyer, 1998; Liebman and Mahoney, 2013; Larkin, 2014).

[5]The general case of intrinsic motivation has been discussed by Tirole (1986) and in later papers (Dewatripont et al., 1999; Akerlof and Kranton, 2005; Besley and Ghatak, 2005; Prendergast, 2007). Physicians balancing profit and patient welfare has been considered by Ellis and McGuire (1986), for example. In contrast, related empirical work has been relatively new, e.g., peer effects due to social incentives (Bandiera et al., 2005, 2009; Mas and Moretti, 2009) and the response to information arguably orthogonal to profits (Kolstad, 2013).

[6]A related result of technological advances is specialized knowledge, which requires care delivered in teams. Although technological advances have been widespread, see Messerli et al. (2005) for the particularly impressive example of modern cardiovascular care, compared to Dwight Eisenhower's heart attack treatment in 1955.

flexible hours (e.g., Goldin, 2014; Presser, 2003). Thus, increasingly, health care is delivered by organizations, and schedules play an important role in assigning uncertain work (e.g., Briscoe, 2006; Casalino et al., 2003). These changes of course have parallels in other industries, which also feature increasingly interrelated and complex production.

The remainder of this paper is organized as follows: Section 2 describes the institutional setting and data. Section 3 discusses a conceptual framework to consider EOS effects. Section 4 investigates physician acceptance of new patients. Section 5 reports EOS effects for patients who are accepted and considers evidence for patient selection and physician fatigue. Section 6 considers the relationship between shift overlap, workload, and patient-care distortion. Section 7 presents simulations of counterfactual regimes of patient assignment. Section 8 discusses additional points of interpretation, and Section 9 concludes.

2 Institutional Setting and Data

2.1 Shift Work

I study a large, academic, tertiary-care ED with a high frequency of patient visits. Like in virtually all other EDs around the country, work is organized by shifts. In the study sample from June 2005 to December 2012, shifts range from seven to twelve hours in length (ℓ). Shifts also differ in overlap with a previous shift ($\underline{o}$) or with a subsequent shift ($\overline{o}$) in the same location. I observe 23,990 shifts in 35 different shift types summarized by $\langle \ell, \underline{o}, \overline{o} \rangle$ (Table A-5.2).

For physicians working in these shifts, the end of shift (EOS) is simply the time after which they are allowed to go home if they have completed their work. Because I focus on behavior at EOS, I pay special attention to $\overline{o}$. This overlap is the time prior to EOS during which a physician shares new work with another physician who has begun work in the same location.[7] "Location" refers to a set of beds in the ED in which a physician may treat patients. This managerial definition may differ from broader physical areas, or "pods," where physicians may see each other but may not share the same beds. That is, a pod may contain more than one managerial location. During my sample period, I observe two to three pods, with a new pod

[7] I distinguish between shifts that end with the closure of a patient location, or "terminal shifts" with $\overline{o} = 0$, and those continuing patient care with another shift in the same location, or "transitioned shifts" with $\overline{o} > 0$.

opening in May 2011, that at various times were divided into two to five managerial locations.

In the study period, the ED underwent 15 different shift schedule changes at the location-week level. Within each regime, the pattern of shifts could differ across day of the week. As is common in scheduled work, shift times were designed around estimated workload needs, and schedule changes reflected changes in the flow of patients to ED. Some shift regime changes were merely minor tweaks in the times of specific shifts, while others involved larger changes.[8] All regime changes, however, can be summarized as a set of shifts, each described by a shift type $\langle \ell, \underline{o}, \overline{o} \rangle$, a starting day and time, location, and range of months that the shift was in effect (see Figure A-5.1; Table A-1 details these shift descriptions).

Shifts are scheduled many months in advance, and physicians are expected to work in all types of shifts at all times and locations. Physicians may only request rare specific shifts off, such as holidays and vacation days, and shift trades are rare. During a shift, physicians cannot control the volume of patients arriving to the ED or the patient types that the triage nurse assigns to beds. Throughout the entire study period, physicians were exposed to the same financial incentives: They were paid a clinical salary based on the number of shifts they work with a 10% productivity bonus based on clinical productivity (measured by Relative Value Units, or RVUs, per hour) and modified by research, teaching, and administrative metrics.[9] Although their salaries are based on numbers of shifts worked, physicians are not compensated for time worked past EOS.[10]

2.2 Patient Care

After arrival at the ED, patients are assigned to a bed by a triage nurse. This assignment determines the managerial location for the patient and therefore the one or more physicians who may assume care for the patient. Once the patient arrives in a bed, a physician may sign

[8] In particular, the regime change in May 2011 included the introduction of a new pod to increase the number of available beds to meet increasing ED volume.

[9] The metric of Relative Value Units (RVUs) per hour is a financial incentive that encourages physicians to work faster, because RVUs are mostly increased on the extensive margin by seeing more patients and are rarely increased by doing more for the same patients.

[10] This is the standard financial arrangement for salaried physicians across the US. Specifically, physicians are exempt from overtime pay as per the Fair Labor Standards Act of 1938 (FLSA). A large number of worker categories are exempt from overtime pay, including most positions with a high degree of discretion (see `http://www.dol.gov/elaws/esa/flsa/screen75.asp`).

up for that patient on the computer order entry system. Physicians are expected to complete work on any patient for whom they have assumed care, in order to reduce information loss with hand-offs (e.g., Apker et al., 2007), except in uncommon cases where the patient is expected to stay much longer in the ED. Because of this, physicians report often staying two to three hours past EOS.[11] For patients arriving near EOS, physicians may opt not to start work and leave the patient for another physician. This option is more acceptable if this physician peer will arrive soon or has already arrived in the same location.

In addition to the attending physician (or simply "physician"), patient care is also provided by resident physicians or physician assistants and by nurses (not to be confused with the triage nurse). These other providers also work in shifts. Generally shifts of different team members do not end at the same time as each other, except when a location closes. More importantly, unlike physicians, care by nurses, residents, and physician assistants is more readily transferred between providers in the same role when they end their respective shifts, perhaps reflecting the lesser importance of their information in decision-making. For example, only physicians have the formal authority to make patient discharge decisions.

For physicians in the ED, the concept of patient discharge is a matter of discretion. Patient care is usually expected to continue after discharge, in either outpatient or inpatient settings. The key criterion for completion of work – or discharge – is whether the physician believes that sufficient information has been gathered for a discharge decision out of the ED. This decision is often made with incomplete diagnosis and treatment. Rather, the physician may decide to discharge a patient home with outpatient follow-up after "ruling out" serious medical conditions, or the physician may admit the patient for inpatient care if the patient could still possibly have a serious condition that would make discharge home unsafe.[12]

Physicians may gather the information they need to make the discharge decision in several ways. Formal diagnostic tests are an obvious way to gain more information on a patient's clinical condition. Treatment can also inform possible diagnoses by patient response, such as response

[11] In shifts with greater overlap, which have become more common, physicians report staying shorter amounts of time, but still up to one hour past EOS. Quantitative evidence using physician orders and patient discharge times is presented in Figure A-5.2 with a brief discussion in Appendix A-5.

[12] In this ED, there is yet a third discharge destination to "ED observation," if the patient meets certain criteria that make discharge either home or to inpatient unclear and justify watching the patient in the ED for a substantial period of time (usually overnight) to watch clinical progress.

to bronchodilators for suspected asthma. But time – for a careful history and physical, serial monitoring, or a well-planned sequence of formal tests and treatment – remains an important input in the production of information. Diagnostic tests and treatments can be complements or substitutes for time: Formal tests (e.g., CT and MRI scans) take time to complete and can thus prolong the length of stay, but testing can also substitute for a careful questioning or serial monitoring to gather information more rapidly.

2.3 Observations and Outcomes

From June 2005 to December 2012, I observe 442,244 raw patient visits to the ED. I combine visit data with detailed timestamped data on physician orders, patient bed locations, and physician schedules to yield a working sample of 372,224 observations. Details of the sample definition process are described in Table A-5.1. In the sample, I observe the identities of 102 physicians, 1,146 residents and physician assistants, and 393 nurses.

Table A-5.2 summarizes the number of observations for each shift type, in terms of hours, potential patients who arrive during a time when a shift of that type is in progress, and actual patients who are seen by a physician working in a shift of that type. Because I focus on behavior near EOS, I also present in Figure 1 key variation across the 23,990 shifts in the time of day for EOS, shift length, and the overlap with another shift at EOS.

ED length of stay not only captures an important input of time in patient care but also largely determines when a physician can leave work. I measure length of stay from the arrival at the pod to entry of the discharge order. The timing of the discharge order, as opposed to actual discharge, is relatively unaffected by downstream events (e.g., inpatient bed availability, patient home transportation, or post-ED clinical care). I also use timestamped orders as measures of utilization and to create intervals of time within length of stay that are likely to be rough substitutes or complements with formal utilization, which I discuss further in Section A-2.

Since the primary product of ED care is the physician's discharge decision, I focus on the decision to admit a patient as a key outcome measure, which has has also received attention as a source of rising system costs (Schuur and Venkatesh, 2012; Forster et al., 2003). I accordingly measure total direct costs, including costs incurred both by formal utilization in the ED and

during a subsequent admission.[13] Finally, I measure thirty-day mortality, occurring in 2% of the sample visits, and return visits to the ED within 14 days ("bounce-backs"), occurring in 7% of the sample (Lerman and Kobernick, 1987). However, these latter outcomes are less strongly influenced by the ED physician and depend on a host of factors outside the ED and hospital system, reducing the precision of their estimated effects.

2.4 Patient Observable Characteristics

When patients arrive at the ED, they are evaluated by a triage nurse and assigned an Emergency Severity Index (ESI), which ranges from 1 to 5, with lower numbers indicating a more severe or urgent case (Tanabe et al., 2004). When the patient is assigned a bed, this information is communicated via a computer interface, together with the patient's last name, age, sex, and "chief complaint" (a phrase that describes why the patient arrived at the ED). I observe all this information displayed to physicians prior to patient acceptance.

In addition, I observe patient characteristics that are usually known (if ever) by physicians only after patient acceptance – insurance status, language, race, zip code of residence, and rich diagnostic information – since physicians do not interact with patients or examine their charts prior to accepting them. I codify the diagnostic information into 30 Elixhauser indicators based on diagnostic ICD-9 codes for comorbidities (e.g., renal disease, cardiac arrhythmias) that have been validated for predicting clinical outcomes using administrative data (Elixhauser et al., 1998). Diagnostic codes of course are also partly determined by patient care.

2.5 Descriptive Evidence

Figure 2 shows a plot of the distribution of visits over arrival time prior to EOS and length of stay. Panel A shows the raw patient visit count in each fifteen-minute bin of arrival time interacted with each fifteen-minute bin of length of stay. Some findings are apparent from these visit plots. First, few patients are seen within the last two hours prior to EOS.[14] Second, lengths of stay are shorter for patients who arrive and are accepted by a physician closer to EOS than for

[13] Direct costs are for services that physicians control and are directly related to patient care. Indirect costs include administrative costs (e.g., paying non-clinical staff, rent, depreciation, and overhead).

[14] Although relatively few patients are also seen arriving greater than nine hours prior to EOS, this fact reflects that relatively few shifts are greater than nine hours in length.

patients farther from EOS. There also appears to be an additional density of visits just prior to the 45-degree line mapping when length of stay roughly equals the time prior to EOS, implying that patients are more likely to be discharged just prior to EOS than at times before or after.

In order to examine more closely the discharge of patients conditional on acceptance, I plot in Panel B of Figure 2 the density of length of stay conditional on arrival time (and acceptance) prior to EOS. This plot shows a greater density of early discharges with arrival times closer to EOS. As in Panel A, for visits with arrival times between two to seven hours prior to EOS, there appears to be a linear mass of discharges along the 45-degree line in which discharges are roughly just prior to EOS.

3 Conceptual Framework

I introduce a simple model to consider how physician decisions – accepting patients and choosing inputs to care – may be distorted under work schedules. While the model is tailored to ED physicians, the key distortionary elements of the model are the following: (1) Workers have private information about their tasks; (2) workers care less about the social consequences of their actions, relative to their own income and leisure; and (3) *ex post* worker-task specificity prevents workers from simply passing off tasks at EOS (Briscoe, 2007; Goldin, 2014).

3.1 Model Setup

Consider a physician in a shiftwork arrangement: She has a contract to arrive at shift beginning $\underline{t}$ and stay until EOS $\bar{t}$ or whenever she discharges her last patient, whichever is later, and she will receive a lump-sum payment y for this. Now consider a patient arriving at time $t < \bar{t}$. The relevant welfare parameters of her work environment is captured by $\mathcal{E}_t \equiv (\mathcal{W}_t, \mathcal{W}'_t)$, where $\mathcal{W}_t \equiv (\underline{t}, w_t)$ includes the start time of the physician's shift, $\underline{t}$, and her current workload, w_t, and $\mathcal{W}'_t \equiv (\underline{t}', w'_t)$ describes similar information for a potential peer in the subsequent shift in the same managerial location. The patient's underlying health state, $\theta \in \{0, 1\}$, is unobservable, but $\Pr\{\theta = 1\} = p$ is publicly known. The physician takes the following actions:

1. Given t, $\mathcal{E}_t$, and p, the physician decides on $a \in \{0,1\}$, whether to accept the patient ($a=1$) or not ($a=0$).

2. If she accepts the patient, she observes private information $\mathcal{I}$ so that $\Pr\{\theta=1|\mathcal{I}\}=p'$, and $|p'-\theta|<|p-\theta|$.[15] She decides on inputs $\mathbf{z}$ in patient care: time τ and formal tests and treatments z.

3. The physician observes θ with probability $q(\mathbf{z}) \in (0,1)$ and decides on $d \in \{0,1\}$, to admit ($d=1$) or discharge home the patient ($d=0$).

4. The patient's health state θ is observed, and the physician receives the following utility:

$$u(t,\mathcal{E}_t;\theta;a,\mathbf{z},d)=\begin{cases} y+\lambda O(\theta;\mathcal{E}_t), & a=0 \\ y-\tilde{c}_\tau(\tau)+\lambda\left(V(\theta,d)-c(\mathbf{z})\right), & a=1 \end{cases}. \quad (1)$$

   Utility is stated in dollar terms, where physician income y does not depend on her actions.[16] $O(\theta;\mathcal{E}_t)$ is the value of the "outside option" if $a=0$, which depends on θ and the work environment $\mathcal{E}_t$. $V(\theta,d)$ is the value of making the right discharge decision. $c(\mathbf{z})$ is the cost of patient care inputs, from which I separate $\tilde{c}_\tau(\tau)$, the cost of foregone leisure if the physician stays past EOS. $\lambda \in (0,1)$, and $1-\lambda$ is the wedge by which the physician undervalues the mission of patient care.

To be clear about the wedge, first consider the social welfare function as equivalent to Equation (1), except without λ (i.e., $\lambda=1$). As $\lambda \to 1$, physician utility approaches social welfare, and the agency problem disappears. As $\lambda \to 0$, utility approaches the standard labor supply model in which workers only care about consumption and leisure. If $\lambda=0$ (which I rule out), the physician would have no incentive to make the right decisions (despite observing $\mathcal{I}$ and sometimes θ).

[15] I rule out private information before patient acceptance in this model. This is generally consistent with the institutional setting, and I examine selection empirically in Section 5.2.

[16] In scheduled work y for the most part depends *ex ante* availability, not *ex post* time past EOS. This model can accommodate some rewards correlated with staying past EOS (e.g., financial incentives for seeing more patients, social recognition); all that it requires is that physicians are relatively uncompensated for leisure.

3.2 Patient Care

I first examine EOS effects on the inputs to patient care and the discharge decision, assuming that the physician has chosen to accept the new patient ($a = 1$). Discharge decisions have important efficiency implications for resource utilization and patient health. Formally, patients with $\theta = 0$ should be discharged home, while those with $\theta = 1$ should be admitted: $V(0,0) > V(0,1)$ and $V(1,1) > V(1,0)$. Discharging a sick patient home is particularly harmful, or equivalently, physicians are risk-averse: $V(1,1) - V(1,0) > V(0,0) - V(0,1)$. Because of this last fact, if θ remains unobserved, the physician will admit if and only if $p' > p^*$, where $p^* < \frac{1}{2}$.[17]

Patient care increases the probability q of observing θ and therefore appropriate discharges.[18] This probability is increased by formal diagnostic tests and treatment, z, and by clinical observation and reasoning over time, τ. q is increasing and concave with respect to τ and z. τ and z may be net substitutes ($\partial^2 q/(\partial\tau\partial z) < 0$) or net complements ($\partial^2 q/(\partial\tau\partial z) > 0$) in production. Effective time per patient is reduced with higher workload w_t: $\partial^2 q/(\partial\tau\partial w_t) < 0$. This contrasts with formal inputs, for which I make the normalizing assumption $\partial^2 q/(\partial z\partial w_t) = 0$.[19] Costs in $c(\mathbf{z})$ and $\tilde{c}_\tau(\tau)$ are positive, continuous, increasing, and convex in their arguments. Define $\tilde{c}_\tau = 0$ for $\tau + t - \bar{t} \leq 0$. For simplicity, assume additive separability of each element of $\mathbf{z}$ in $c(\mathbf{z})$.

Proposition 1. *Denote decisions in Section 3.1 that maximize expected utility in Equation (1), conditional on patient acceptance ($a = 1$), as τ^*, z^*, and d^*. Denote corresponding decisions that maximize welfare as τ^{FB}, z^{FB}, and d^{FB}.*

(a) As $t \to \bar{t}$, τ^* weakly decreases, z^* may weakly increase (if τ and z are net substitutes) or decrease (if τ and z are net complements), and $E[d^*]$ weakly increases as long as $F_{p'}(p^*) < \frac{1}{2}$.

(b) For all t, $\tau^* < \tau^{FB}$, and $E[d^*] < E\left[d^{FB}\right]$.

(c) If τ and z are net substitutes, then $z^* > z^{FB}$, and $z^* - z^{FB}$ weakly increases in w_t,

[17] This can be straightforwardly shown by noting that $E[V|d = 0, p' = p^*] = E[V|d = 1, p' = p^*]$.

[18] I abstract away from treatment within the ED that can improve the patient's health. This can easily be incorporated into the model and would not change qualitative results, except that if z^* is increasing in p, then physicians will be less likely to accept *ex ante* sicker patients.

[19] The intuition behind this is that with more patients, a physician has to divide her time and attention between them, but formal utilization can be ordered with the click of a mouse. Any additional time implications (e.g., time for initial evalaution or to review CT scans) would be incorporated in τ. By the normalizing assumption, I focus attention on substitutability or complementarity between time and formal utilization.

holding t constant and for all t. The reverse is true if τ and z are net complements dominates.

As the physician nears EOS, she will shorten length of stay τ. The intensity of diagnostic tests and treatments may increase or decrease, depending on whether τ and z are net substitutes or complements, respectively. Finally, she observes θ with lower probability q. This increases admissions, as long as $F_{p'}(p^*) < \frac{1}{2}$, where $F_{p'}(\cdot)$ is the c.d.f. of p' conditional on p and $a^* = 1$ (i.e., as long as $\theta = 1$ with sufficient probability). These distortions increase with workload w_t, which further increases the cost of time by reducing the effective time per patient to produce q.

3.3 Patient Assignment

I next consider the physician's upstream decision to accept the new patient, $a \in \{0, 1\}$. The physician compares the outside option under $a = 0$, including whether the patient is likely to wait for care, and expected utility under $a = 1$,

$$E\left[u\left(t, \mathcal{E}_t; \theta; 1, \mathbf{z}^*, d^*\right)\right] = y + \max_{\mathbf{z}} \left\{ \lambda \left(E\left[\max_{d} V\left(\theta, d\right) \right] - c\left(\mathbf{z}\right) \right) - \tilde{c}_\tau\left(\tau\right) \right\},$$

where

$$E\left[\max_{d} V\left(\theta, d\right)\right] = \begin{cases} E\left[V\left(\theta, 0\right)\right] + pq\left(V\left(1,1\right) - V\left(1,0\right)\right), & p < p^* \\ E\left[V\left(\theta, 1\right)\right] + \left(1-p\right)q\left(V\left(0,0\right) - V\left(0,1\right)\right), & p \geq p^* \end{cases}.$$

Denote $\underline{O}^*$ as the threshold rules such that accepting the patient maximizes expected utility ($a^* = 1$) if and only if $E\left[O\left(\theta; \mathcal{E}_t\right)\right] > \underline{O}^*$. It is easy to see that $\underline{O}^* = W\left(\mathbf{z}^*, d^*\right) - \left(\lambda^{-1} - 1\right)\tilde{c}_\tau\left(\tau^*\right)$, where $W\left(\mathbf{z}, d\right) \equiv E\left[V\left(\theta, d\right)\right] - c\left(\mathbf{z}\right) - \tilde{c}_\tau\left(\tau\right)$. The corresponding first-best threshold that determines the first-best acceptance a^{FB} is $\underline{O}^{FB} = W\left(\mathbf{z}^{FB}, d^{FB}\right)$, when optimal $\mathbf{z}$ and d can be implemented. Finally, consider the second-best assignment policy, in which the patient may be assigned as a policy, $a^{SB} \in \{0, 1\}$, but the physician controls $\mathbf{z}$ and d. In this policy, $a^{SB} = 1$ if and only $E\left[O\left(\theta; \mathcal{E}_t\right)\right] > \underline{O}^{SB} = W\left(\mathbf{z}^*, d^*\right)$.

Proposition 2. *Consider a^* as the patient acceptance decision in Section 3.1 that maximizes expected utility in Equation (1), a^{FB} as the assignment that maximizes expected welfare when optimal $\mathbf{z}$ and d are publicly known and contractible, and a^{SB} as the assignment that maximizes expected welfare when optimal $\mathbf{z}$ and d are either publicly unknown or non-contractible. Assignment will follow threshold rules in which assignment occurs if and only if $E\left[O\left(\theta;\mathcal{E}_t\right)\right]$ is greater than a threshold. The respective threshold rules are $\underline{O}^*$, $\underline{O}^{FB}$, and $\underline{O}^{SB}$, where $\underline{O}^* < \underline{O}^{SB} < \underline{O}^{FB}$. $\underline{O}^{FB} - \underline{O}^{SB}$ and $\underline{O}^{SB} - \underline{O}^*$ increase as $t \to \bar{t}$ decreases or as λ decreases.*

There are first-best reasons for assignment to decrease near EOS. As $t \to \bar{t}$, the outside option $O\left(\theta;\mathcal{E}_t\right)$ increases because a peer is more likely to be arriving soon or already present, and $W\left(\mathbf{z},d\right)$, holding $\mathbf{z}$ and d fixed, may also decrease due to fatigue and the possibility of foregone leisure.[20] However, beyond this decrease, patient acceptance a^* will be inefficiently low near EOS ($\underline{O}^* < \underline{O}^{FB}$). The second-best policy, in which physicians continue to choose $\mathbf{z}^*$ and d, will assign patients at a threshold $\underline{O}^{SB}$ in between $\underline{O}^*$ and $\underline{O}^{FB}$. The relative distance between these policy thresholds will depend on the curvature of W (i.e., $\left|d^2W/d\tau^2\right|$): If W is not very curved, then patient-care distortions, $W\left(\mathbf{z}^{FB},d^{FB}\right) - W\left(\mathbf{z}^*,d^*\right)$, will be greater relative to the misvaluation of leisure, $\left(\lambda^{-1}-1\right)\tilde{c}_\tau\left(\tau^*\right)$. Thus, $\underline{O}^{SB}$ will be closer to $\underline{O}^*$ than to $\underline{O}^{FB}$.

3.4 Remarks

The inefficiency in the model is fundamentally informational. First, physicians observe private information p', so management does not know $\mathbf{z}^{FB}$ and d^{FB}. Second, they are imperfect agents, overvaluing consumption and leisure relative to patient care. The canonical way to implement first-best would be to pay physicians an hourly overtime wage, in this case $(1-\lambda)\,\tilde{c}'_\tau$.[21] However, this is impractical due to uncertainty and complexity, discussed in Weitzman (1974),

[20] Another version of the patient acceptance question is patient selection (i.e., how $E\left[p\middle|\, a^*=1\right]$ changes as $t \to \bar{t}$). Selection will likely be towards healthier patients: For low p and as $t \to \bar{t}$, expected utility under $a=1$ likely diminishes less quickly, and expected utility under $a=0$ likely increases less quickly. I will examine this empirically in Section 5.2.

[21] This is less than the full marginal cost of labor because of the "compensating differential" utility physicians gain from treating patients.

leading firms to specify schedules and assign work. In fact, prespecifying schedules and pay removes patient-care distortion *within* the shift.[22]

Implicit in this model is a cost that precludes physicians from passing off patients to another physician at EOS or at any other point before patient discharge. With no transfer cost, there would be no EOS distortion. Part of this transfer cost represents a loss of information (e.g., reducing $q(\mathbf{z})$) (Briscoe, 2006, 2007; Goldin, 2014), whereas another part may be due to social distortions (e.g., the desire not give peers work). Patients are rarely transferred in this institutional setting, and I do not observe the exact time of pass-off for the few who are transferred. However, in Section 7, I will empirically assess lower bounds to transfer costs given observed increases in resource-utilization costs.

In this informational environment, work assignment is a natural policy lever, since assignment is easy to observe and influence. Physicians may be assigned too little work to justify the value of their time. However, assigning more work worsens distortions in patient-care decisions, $\mathbf{z}$ and d, both by assigning patients to physicians under time pressure, and via the dynamic of increasing workload and therefore reducing effective time per patient.[23] In Section 6.2, I empirically focus on shift overlap $\overline{o} \equiv \overline{t} - \underline{t}'$ as one mechanism that influences a_t^* through changing $O(\theta; \mathcal{E}_t)$. More broadly, a_t may be implemented by a variety of managerial instruments, such as piece-rate pay, social norms, or formal assignment policies. Therefore, while I model a_t as a physician choice here, in Section 7, I more generally consider it as a sufficient-statistic policy instrument.

4 Patient Assignment

In this section, I describe patient assignment near EOS. As in Proposition 2, it is natural that physicians will be less likely to accept patients as EOS nears, because time for patient care is more costly. Patient assignment to physicians can also be influenced by assignment to locations (particularly in location-times with only one physician). The simple analysis in this

[22] This is possible as long as physicians can be guaranteed to leave by EOS, which can be mostly implemented by avoiding new work near EOS, rather than discharging patients earlier within the shift.

[23] I have not explicitly modeled this dynamic. This may be formally considered in an expanded dynamic model with two patients arriving at different times, t and $t+1$, and respective decisions $(a_t, \mathbf{z}_t, d_t)$ and $(a_{t+1}, \mathbf{z}_{t+1}, d_{t+1})$. Increasing a_t increases w_{t+1} and thus, from Proposition 1, reduces welfare by worsening distortions in $\mathbf{z}_{t+1}^*$ and d_{t+1}^*.

section presents unadjusted average rate of patient assignment to a physician nearing EOS across a variety of shift types. In particular, I will verify that greater overlap $\overline{o}$ allows physicians to decline patients earlier relative to EOS.

Figure 3 presents the hourly average rates of new patient visits, with each panel representing shifts with a different $\overline{o}$, for the index physician (patients accepted), for the location inclusive of the index physician (patients assigned by the triage nurse), and for the entire ED (patients arriving at the ED). Regardless of the shift type, physicians generally accept between two to three new patients per hour at most, and rates of acceptance are highest near the beginning of shift. Thereafter, in transitioned shifts with $\overline{o} > 0$, the average rates of patient flow show two consistent relationships with time. First, patient flow declines precipitously in the hour *prior* to the transitioning peer's arrival at the location. Second, patient flow declines close to zero in the two to three hours prior to EOS. If there is sufficient $\overline{o}$, patient flow is relatively constant but diminished in that duration. In terminal shifts, where $\overline{o} = 0$, the decline in patient flow begins earlier, at least four hours prior to EOS.

Also in Figure 3, patients who are not accepted by the index physician may wait up to an hour to be seen by a peer yet to arrive, but patient flow to transitioning peers generally at least makes up for the decline in flow for the index physician. That is, despite declines in patient acceptance, patients continue to arrive at the pod at similar or greater rates prior to the peer's transitioning shift. Finally, Figure 3 plots the flow of patients to the entire ED, showing background patient flow to other pods that seems unrelated to flows to the index physician. Naturally, overall ED flow appears more stable when averaged across greater shift observations and variation across times of the day (see Figure 1, e.g., $\overline{o} = 1$ and $\overline{o} = 6$).

These relationships are remarkably consistent, over different $\overline{o}$, despite being presented as unadjusted averages. It is intuitive that physicians would decrease their acceptance of new patients as they approach EOS, since the cost of seeing new patients increases with proximity to EOS. The cost is both in the time cost to the physician ending her shift and also in terms of the resulting distortion in patient care.

The earlier arrival of peers allows for earlier reductions in patient assignment relative to EOS. This includes reductions *prior* to peer arrival, especially in shifts with shorter transitions,

suggesting anticipatory behavior. For terminal shifts with no peer arriving in the same location, remarkably, the long decline in patient flow rates is implemented by the *triage nurse* assigning fewer patients to the physician nearing EOS. Thus, "slacking off" is achieved between coworkers sharing a location and, in cases without coworkers, by managerial assignment itself.

5 Effect on Patient Care

5.1 Main EOS Effects

My main analysis addresses the following: What is the effect of a patient's arrival near a physician's EOS on that patient's care by that physician? Although I address patient selection more directly later, I first control for a rich set of patient characteristics. I use variation within the same health care providers working at different times and locations to control for fixed provider unobservables. Using shift variation within locations and within times, I control for unobservables (e.g., patient characteristics and ED resources) that vary by location and time categories, such as time of the day or day of the week. I finally use variation in shift lengths to control for fatigue, which I consider due to time relative to the beginning of shifts.[24]

In the full specification, I estimate the following equation:

$$\begin{aligned} Y_{ijkpt} \quad = \quad & \sum_{m=-6}^{-1} \alpha_m \mathbf{1}\left(\left\lfloor t - \overline{t}\left(j,t\right)\right\rfloor = m\right) + \sum_{m} \gamma_m \mathbf{1}\left(\left\lfloor t - \underline{t}\left(j,t\right)\right\rfloor = m\right) + \\ & \mathbf{X}_{it}'\beta + \mathbf{T}_t'\eta + \zeta_p + \nu_{jk} + \varepsilon_{ijkpt}, \end{aligned} \tag{2}$$

where outcome Y_{ijkpt} is indexed for patient i, physician j (in shift from $\underline{t}\left(j,t\right)$ to $\overline{t}\left(j,t\right)$), assisting team k (including the resident or physician assistant, and the nurse), pod p, and arrival time t. The coefficients of interest in Equation (2) are $\{\alpha_m\}$, or the effect of arrival m hours (rounded down to the nearest negative integer) prior to EOS. I control for time relative to the shift beginning $(t - \underline{t}\left(j,t\right))$, patient characteristics $\mathbf{X}_{it}$, time categories $\mathbf{T}_t$ (for month-year, day of the week, and hour of the day), pod identities ζ_p, and physician-team identities ν_{jk}.

Table 1 shows results for log length of stay, estimating coefficients $\{\alpha_m\}$ for time prior to

[24] In alternative models, I also control for cubic splines of total number of patients seen prior to the index patient's arrival. Results (not shown) are essentially identical with these additional controls.

EOS, from versions of Equation (2) with varying sets of controls. All models estimate highly significant and negative coefficients for approaching time to EOS, with visits seven or more hours prior to EOS being the reference category. The reduction in length of stay grows larger in magnitude as time approaches EOS. By the last hour prior to EOS, versions of Equation (2) estimate effects on log length of stay ranging from -0.53 to -0.72. The full model, shown in the last column of Table 1 and plotted in Panel A of Figure 4, estimates an effect on log length of stay of -0.59 in the last hour and serves as the baseline model for this paper.

The difference in estimates between the first and second columns in Table 1 reflects the change in the estimated effect due to including a rich set of patient characteristics, which is about 0.06 on log length of stay in the last hour prior to EOS. I explore selection more directly below. The difference between the fourth and last columns represents the effect of time relative to shift beginning, which can include fatigue and is separately identified from EOS effects due to variation in shift lengths. This difference, about 0.13 in the last hour prior to EOS, also accounts for only a minor portion of the overall effect.[25]

Table 2 shows results for other outcome measures, including the order count, inpatient admission, log total cost, 30-day mortality, and 14-day bounce-backs. Estimates for α_m are generally insignificant for hours before the last hour prior to EOS, but are significantly positive in the last hour. Patients arriving and accepted in the last hour prior to EOS have 1.4 additional orders for formal tests and treatment, from a sample mean of 13.5 orders.[26] These patients are also 5.7 percentage points more likely to be admitted, which is 21% relatively higher than the sample mean of 27%. Log total costs are 0.21 greater in the last hour prior to EOS. Mortality and bounce-backs do not exhibit a significant effect with respect to EOS, although these outcomes are either rare (mortality) or imprecisely predicted (bounce-backs). I plot coefficients for orders, admissions, and total costs in Panels B to D of Figure 4.

5.2 Patient Selection

Physicians may accept or be assigned healthier patients as they approach EOS. However,

[25] See Appendix A-1 for more direct results on effects relative to shift beginning.

[26] This suggests that formal orders are a net substitute for time. See Appendix A-2 for more direct results supporting this hypothesis.

there are reasons why selection, especially on unobservables, is likely to be limited. Physicians have little scope for selecting patients by characteristics unobservable in the data because norms discourage them from looking behind curtains before choosing patients and thus usually only observe a patient's key descriptors on the computer interface prior to this decision. Furthermore, there are no formal policies (which can be gamed) against engaging in selection, but there are strong norms against such behavior between physicians in the same pod, who likely observe the same information prior to acceptance. Finally, reducing the acceptance rate near EOS is an explicitly tolerated policy, shown in all types of shifts (Section 4). In this section, I empirically assess the extent of selection with four sets of evidence.

First, I summarize observable characteristics of accepted patients by arrival time relative to EOS. In Figures A-3.1 to A-3.5 (details in Appendix A-3.1), mean observable characteristics, such as age, ESI, race, and language, are stable and only slightly trending towards healthier patients as arrival time of the accepted patients nears EOS. Observable EOS selection appears slightly stronger in terminal shifts, in which all selection is due to triage nurse assignment, than in shifts with overlap when physicians choose patients vis-a-vis a peer. Quantiles are also highly stable and show no change in the (large) variation of patients characteristics with arrival time relative to EOS (Figures A-3.6 and A-3.7).

Second, in Appendix A-3.2, I make use of patient characteristics generally unobservable at the time of patient acceptance, such as *ex post* diagnoses or insurance status, in a regression framework to quantify the degree of selection on unobservables. Using characteristics that are generally observed before acceptance ($\mathbf{X}_{it}^{prior}$) and the full set that includes characteristics generally only observed after acceptance ($\mathbf{X}_{it}^{full}$), I form two predicted outcomes, $\hat{Y}_{ijkpt}^{prior}$ and $\hat{Y}_{ijkpt}^{full}$, respectively. I regress these predicted outcomes on arrival time prior to EOS as a method to quantify the degree of selection on observables and the incremental degree of selection on unobservables for each outcome. I find relatively small selection on observables in the direction that predicts shorter lengths of stay near EOS (5.4% shorter in the last hour), but *lower* orders, admissions, and costs, the opposite of what I find for these latter outcomes. More importantly, incremental selection on unobservables is essentially nonexistent.

Third, in Appendix A-3.3, I undertake an analysis, based on Altonji et al. (2005), to compute

the degree of selection on patient unobservables relative to selection on observables that would be required to explain my length of stay results. This approach considers, for patients arriving at each hour prior to EOS, the explanatory power of observables in determining whether these patients are accepted and the explanatory power of observables in determining length of stay. I find that selection on patient unobservables must be 475 times greater than selection on observables in order to explain the entire effect on length of stay for patients arriving in the last hour prior to EOS.

Fourth, in Appendix A-3.4, similar to an approach taken by Chetty et al. (2014), I only use variation in the overall set of shifts in progress at a given hour for the *entire ED.* Averaging patients within hour of arrival eliminates the potential bias due to unobserved selection across physicians. I therefore compare predictions based on estimated EOS effects with actual residual log length of stay, averaging both over patients within each hour, in order to estimate bias due to selection across physicians within hour. In Panel A of Figure 6, the ED shift environment predicts average actual length of stay, with no evidence of bias: The relationship between the shift-environment prediction and actual log length of stay is linear with a slope of 1.029 (t-value of 17.16). In contrast, in Panels B and C, the ED shift environment is unrelated to length of stay predicted by $\mathbf{X}_{it}^{prior}$ or $\mathbf{X}_{it}^{full}$, suggesting that the arrival times of patients differing by (observable) types are not correlated with the ED shift environment.

6 Shift Overlap, Workload, and Distortion

I evaluate how workload and patient-care effects vary across shifts with varying overlap near EOS, for two purposes: First, this supports the interpretation that EOS effects reflect inefficiency, under the identifying assumption that the EOS by itself has no first-best implications for patient care, conditional on volume of work, time since beginning work, and time since a peer's arrival. Formal overlap (i.e., time between peer arrival and EOS) only changes when a physician is *allowed* to leave work. Second, this analysis uses shift structure as a concrete example of patient assignment as a policy lever. Through patient assignment, a planner can influence the efficiency of patient care: Assigning fewer patients to physicians on schedules may underutilize the value of their time, but assigning more patients worsens the EOS distortion in

the use of time as an input to care.

6.1 Patient Censuses over Time

As a descriptive exercise, I first measure workload w_{jt} as the number of patients cared for by physician j (her "census") at time t:

$$w_{jt} = \sum_{J(i,t')=j} \mathbf{1}\left(t \geq t'\right) \mathbf{1}\left(t \leq t' + \tau\left(i, t'\right)\right), \tag{3}$$

the patients accepted at $t' \leq t$ and had length of stay $\tau\left(i, t'\right)$ such that $t \leq t' + \tau\left(i, t'\right)$, where $J\left(i, t'\right)$ is a function assigning patient i arriving at t' to a physician.

Figure A-5.3 shows unadjusted census averages in 30-minute intervals in different shift types by $\overline{o}$. On average, censuses start at around two patients at the beginning of all shift types, representing unstaffed patients from the previous shift, except for shift types with $\overline{o} = 2$, which happen not to transition from another shift. Patients remain on the census at EOS. The number of patients remaining on census in the last 30 minutes prior to EOS is consistently close to four, with the exception of shifts with $\overline{o} = 1$, which have censuses of about six.

Because of client-worker specificity, physicians must usually either reduce censuses to close to zero or at least have a well-defined pass-off plan to instruct another physician.[27] With smaller $\overline{o}$, physicians are *allowed* to go home (at EOS) at an earlier point relative to peer arrival. Holding constant time from beginning of shift and time from peer arrival, smaller $\overline{o}$ thus induces a greater scope for distortionary care near EOS.

6.2 EOS Effects by Shift Overlap

I then consider how patient-care EOS effects may differ by shift overlap. Larger patient-care effects with small $\overline{o}$, conditional on time from beginning of shift, are consistent with distortionary care. Further, the interaction provides evidence of the intuitive tradeoff between extensive and intensive margins of distortion: If physicians have more time to slack off before EOS, workload

[27] As described in Section 2.2, physicians report that they therefore generally stay at least an hour after EOS. This is also supported by the timing of physician orders relative to EOS, shown in Figure A-5.2. Also, while physicians have an average of four patients at EOS in shifts with $\overline{o} = 0$, these patients are much harder to transfer than in transitioned shifts.

near EOS will be lower, and patient-care distortions will be smaller.

I consider three categories of overlap at EOS – terminal shifts ($\overline{o}=0$), minimally transitioned shifts ($\overline{o}=1$), and substantially transitioned shifts ($\overline{o}\geq 2$)[28] – and estimate

$$\begin{aligned} Y_{ijkpt} &= \sum_{m=-6}^{-1}\sum_{\overline{O}}\alpha_{ms}\mathbf{1}\left(\left\lfloor t_i-\overline{t}\left(j,t\right)\right\rfloor=m\right)\mathbf{1}\left(\overline{o}\left(j,t\right)\in\overline{O}\right)+ \\ &\quad \sum_{m}\gamma_m\mathbf{1}\left(\left\lfloor t-\underline{t}\left(j,t\right)\right\rfloor=m\right)+\mathbf{X}_{it}'\beta+\mathbf{T}_t'\eta+\zeta_p+\nu_{jk}+\varepsilon_{ijkpt}, \end{aligned} \tag{4}$$

similar to Equation (2) but interacting the hourly EOS effects by overlap $\overline{o}\left(j,t\right)$ in categories $\overline{O}$. I normalize coefficients so that, as before, the reference category includes times seven hours or greater prior to EOS in each of the overlap categories.

Figure 7 shows EOS effects, across the three categories of shift types, for length of stay, orders, admission, and total costs. The EOS effect on length of stay is largely similar among shift categories (Panel A). All three shift categories show a substantial decline in length of stay as EOS approaches. However, EOS effects are notably absent in shifts with $\overline{o}\geq 2$ for orders, admission probability, and total costs (Panels B to D). In contrast, shifts with $\overline{o}\leq 1$ show large increases in orders, admissions, and total costs at EOS.

6.3 Effective Time per Patient

The evidence above supports highlights a link between patient assignment, workload, and patient care: Assigning physicians more patients near EOS increases workload and thus decreases the effective time physicians spend on each patient's care. In order to operationalize this concept, I create a new outcome measure of workload-adjusted length of stay, which normalizes length of stay by the physician's average census during the stay. That is, for patient i accepted at time t, I divide length of stay (τ_{it}) by the average census under physician $J\left(i,t\right)$ during the i's length of stay ($\overline{w}_{it}$):

$$\tau_{it}/\overline{w}_{it}=\tau_{it}\left[\frac{1}{\tau_{it}}\int_{\tilde{t}\in[t,t+\tau_{it}]}w_{J(i,t),\tilde{t}}d\tilde{t}\right]^{-1}, \tag{5}$$

[28] While I observe shifts with $\overline{o}\in\{2,3\}$, they entail very few observations, as listed in Table A-5.2. Results are essentially unchanged whether I omit these observations or consider them as belonging to the minimally transitioned shift category.

where census w_{jt} is defined by Equation (3).

I then regress the log of workload-adjusted length of stay using Equation (2).[29] As shown in the last column of Table 2, time relative to EOS has little effect on workload-adjusted length of stay until the last hour prior to EOS, when this measure decreases significantly. Thus, adjusting length of stay for workload reconciles previous results in which length of stay progressively decreases as EOS approaches, but orders, admissions, and costs increase only in the last hour. At least in sample, distortions in patient care, including the use of time, appear to only become significant in the last hour prior to EOS. Similarly, in Table 3, I consider effects on workload-adjusted length of stay by shift overlap and find that it also decreases substantially only in the last hour prior to EOS when $\overline{o} \leq 1$. In contrast, when $\overline{o} \geq 2$, workload-adjusted length of stay does not decrease near EOS and, if anything, slightly increases prior to the last hour of shift.[30]

7 Counterfactual Assignment Policies

Despite important variation in patient assignment across shifts with different $\overline{o}$, observed assignment – either between physician peers or by the triage nurse – dramatically diminishes near EOS in all shifts. In this section, I consider the assignment of work as a sufficient statistic for a wide range of managerial policies (e.g., rules, financial incentives) including but not limited to shift overlap. Using a model of patient assignment, discharges, workload, and cost, I assess the efficiency implications of a fuller range of counterfactual assignment policies.

The intuition is that, while work assignment is easily observable and therefore a natural managerial policy, the downstream effects on patient care – particularly the use of time – are much more difficult to monitor or manage. I therefore allow physicians full discretion in how they respond to counterfactual assignment policies, and I empirically calibrate their behavior to match data observed over the range of shift overlap. To evaluate welfare, I consider overall costs due to physician time, patient time, and hospital resources. While assigning more patients

[29] This is different than controlling for current census; results in Table 1 are unchanged when flexible splines of current census are included in Equation (2). Instead, workload-adjusted length of stay solely captures *future* actions by the physician, including future censuses. Otherwise including future censuses as covariates in a regression framework would be problematic.

[30] Such potential increases in workload-adjusted length of stay above baseline do not appear to be associated with increases or decreases in other outcomes of orders, admissions, or costs. This could be consistent with *increases* in length of stay for strategic purposes, or "foot-dragging," as discussed in Chan (2015).

to physicians near EOS mechanically reduces the number of physician-hours (i.e., the amount of overlap) to process patient flow, it worsens other resource costs by increasing workload when there is distortionary pressure to leave work.

7.1 Simulation Routine

To calibrate this model with the data, I estimate discrete-time functions for patient assignment and discharge that crucially depend on time to EOS. Patient discharge follows a hazard model $\mathcal{D}(t, \sigma_s, w_{jt}, \hat{\tau}_{ist})$ that depends on time t, shift characteristics σ_s for shift s (i.e., shift type $\langle \ell, \underline{o}, \overline{o} \rangle_s$ and time of EOS $\overline{t}(s)$), physician j's workload w_{jt} at t, and patient i's predicted length of stay $\hat{\tau}_{ist}$ (details are given in Appendix A-4, including model fit, shown in Figure A-5.4). Assignment $\mathcal{A}(t, \sigma_s, w_{j,t-1})$ follows a zero-inflated Poisson process that similarly depends on time t, shift characteristics σ_s, and workload $w_{j,t-1}$ from the previous period. Although realized assignment is stochastic, I take the *ex ante* assignment policy as under the control of a planner.[31] The patient-assignment function allows a convenient specification of counterfactual policies in which assignment is modified only by how time to EOS is considered. For example, a counterfactual assignment policy may assign more patients one hour prior to EOS by assigning as if the time were three hours prior to EOS. Although patient assignment may be modified by policy, physicians continue to discharge patients with their own discretion. The discharge decision reflects not only how time is used but also, through the relationship between workload-adjusted length of stay and costs, determines additional costs that derive from the EOS distortion.

Specifically, I parameterize counterfactual assignment policies

$$\mathcal{A}_{\Delta}(t, \sigma_s, w_{j,t-1}) \equiv \mathcal{A}\left(\breve{t}(t, s, \Delta), \sigma_s, w_{j,t-1}\right),$$

where the index Δ represents a time shift in observed assignment patterns near EOS. Intuitively, if $\Delta < 0$, assignments are "curtailed," using $\breve{t} > t$ in the assignment function, by at most $|\Delta|$ hours earlier before EOS. If $\Delta > 0$, assignments are "extended" by at most Δ hours, using $\breve{t} < t$. Equation (A-4.17) describing $\breve{t}(t, s, \Delta)$ and other details are in Appendix A-4. Figure 8 shows

[31] Of course, an "assignment policy" could result endogenously from physicians responding to shift structure, as in the conceptual framework (Section 3), but I do not model this intermediate step and simply assume that any assignment policy can be implemented.

example counterfactual policies, for $\Delta \in \{-4, -2, 2, 4\}$.

For each counterfactual policy $\Delta \in [-4, 4]$, I simulate assignments and discharges using functions $\mathcal{A}_\Delta(t, \sigma_s, w_{j,t-1})$ and $\mathcal{D}(t, \sigma_s, w_{jt}, \tilde{\tau}_{ist})$, respectively, where $\tilde{\tau}_{ist}$ is a further prediction of $\hat{\tau}_{ist}$ to simplify computation. In each simulation $r = 1, \ldots, 100$ of each policy Δ, I calculate total counterfactual costs,

$$\text{Costs}^r_\Delta = \text{PhysicianTime}^r_\Delta + \text{PatientTime}^r_\Delta + \text{HospitalResources}^r_\Delta, \tag{6}$$

which I take as a measure of welfare, under the conservative assumption that patient health is unaffected despite EOS distortions in time, formal utilization, and admissions.[32]

$\text{PhysicianTime}^r_\Delta$ captures additional wages that the ED must pay in order to meet patient flow. Physician-time costs may increase for two reasons. First, if fewer patients are scheduled prior to EOS, a peer must arrive earlier because backlog occurs earlier. This is mechanically related to assignment. Second, if more patients are assigned prior to EOS, the index physician must stay later past EOS, and this foregone leisure is valuable. This is not only related to assignment, but also to physician discharge responses to patient load and time relative to EOS. To value physicians time, I use a base-case wage of \$120/hour, which is close to actual wages in this ED and national averages of hourly pay, although results are insensitive to wages several multiples higher. I also value patient time, in $\text{PatientTimeCosts}^r_\Delta$, a \$20/hour, so that shorter lengths of stay are valuable from a patient-perspective, all else equal.

$\text{HospitalResources}^r_\Delta$ captures changes in formal utilization and admissions as physicians spend less time on patients. Based on evidence in Section 6.2, I consider changes in these costs as distortions. Specifically, in each simulation of an assignment policy, I measure decreases in workload-adjusted length of stay near EOS and then simulate increases in per-patient costs, using a calibrated cost elasticity of -1.15 in response to workload-adjusted length of stay.[33] As

[32] In sample, recall that I find no effect on mortality or bounce-backs (Table 2), although this may not hold out of sample. However, this should not matter for the optimal assignment if the optimal assignment policy occurs close to the observed assignment regime, which I show below.

[33] The elasticity estimate is motivated by the fact that both observed total costs (Figure 4) and observed workload-adjusted length of stay (Figure A-5.5) increase only in the last hour prior to EOS, in Section 6.2. In simulated data, I calculate workload-adjusted length of stay decreases by 18.1% in the last hour of shift when $\Delta = 0$, an estimate very close to but more conservative than based on actual data in Table A-5.3. Since total costs increase by 20.8% in the last hour prior to EOS, I calculate the elasticity as $20.8\%/-18.1\% = -1.15$. More

more patients are assigned near EOS, resource costs are distorted upwards, on a per-patient basis and applied to more patients. These costs are empirically based on total direct costs in the hospital accounting data, which reflect the value of resources such as nursing time, tests, treatment, and ED and hospital bed availability.

7.2 Results

Increasing assignment near EOS results in a large increase in resource-utilization costs, which dominates any physician-time savings that may accrue from a later-arriving peer. For example, an assignment policy that results in physicians staying an extra hour past EOS also induces them to spend an extra $5,500 in resource-utilization costs per shift. Figure 9 shows average changes in total costs per shift, stated in Equation (6), under counterfactual policies, where policies are shown in terms of changes in the number of patients assigned.[34] Both curtailing and extending patient assignment increase overall costs relative to those under the actual assignment policy, suggesting that the observed pattern of assignment is approximately second-best optimal. This is true even under an extreme assumption of a $600 per hour physician wage.

Another way to use this simulation is to assess the implicit tradeoff physicians make between foregoing leisure and increasing resource-utilization costs. At each point in time relative to normal completion, I compute the dollar value of extra resource-utilization costs incurred per leisure hour gained, shown in Figure 10 (details in Appendix A-4.4). As indicated earlier, the actual assignment policy results in minimal patient-care distortions, reflecting a low "value" of leisure (below the market wage of $120 per hour) prior to actual completion of work, even though normal completion is reportedly two to three hours past EOS. This could reflect norms to stay past EOS, which I discuss in the subsequent section. However, the value of leisure quickly rises above market wage at 15 minutes past the normal time of completion. By one hour past this time, physicians are willing to expend $990 in order to avoid an additional hour at work. Under a strict interpretation of Equation (1), this implies $\lambda = \$120/\$990 = 0.12$.[35]

detail is given in Appendix A-4.3.

[34] Changes in costs with respect to changes in patients assigned is easier to understand than the policy index Δ, since Δ is only *maximum* amount of change in time in the counterfactual assignment policy (i.e., $\left|\check{t} - t\right| \leq \Delta$). This can be appreciated in plots of curtailed assignment policies in Figure 8, in which $\Delta = -2$ is still very similar to $\Delta = 0$.

[35] Recall that I do not estimate a single λ during calibration. Rather, the simulation exercise calibrates the

Finally, this simulation may shed light on the cost of transferring patients to another physician. Without this transfer cost, there would be no EOS distortion. While transferring patients is uncommon, and while I do not explicitly model patient transfers after EOS, one way to view the transfer cost is that it must at least as large as the additional cost induced by assigning patients near EOS.[36] Otherwise, physicians could reduce costs by transferring the additional patients. I therefore estimate lower bounds on per-patient transfer costs, in assignment regimes where more patients are assigned, by dividing increases in hospital-resource costs by increases in patients assigned. As shown in Figure 11, these bounds are significant, ranging from $220 to $1250 per patient, or 25% to 140% of the base hospital-resource costs.[37]

8 Discussion

The main focus of this paper is to assess a simple but, to my knowledge, unexplored consequence of work schedules: When work past scheduled availability is undercompensated, workers will avoid new work, and the use of time for work will be distorted, possibly in costly ways. While these effects are illustrated concretely in the setting of health care, there are several general points of interpretation. I discuss some of these briefly here.

Presenteeism and Slacking Off. The terms "presenteeism" and "slacking off" have become common in everyday usage. Some definitions of presenteeism describe workers "stay[ing] beyond the time needed for effective performance on the job" (Simpson, 1998).[38] Slacking off has been described as tapering work, particularly in the context of shirking near the end of scheduled work.[39] Both of these concepts are related to the phenomenon described in this paper. Despite

physician discharge decision to many moments in the data, e.g., average censuses and lengths of stay during each 30-minute interval for each shift type (see Figure A-5.4). Given no (assumed) worsening of patient health, these values of extra resource-utilization costs per leisure hour gained thus reveals λ at each point in time.

[36]This exercise thus states the transfer cost in terms of dollars of hospital resources. As mentioned in Section 3, however, the transfer cost includes both patient-care concerns (denominated in hospital dollars) and social costs of violating norms or giving peers extra work (denominated in income dollars).

[37]These lower-bound estimates are monotonic with increasing assignment, which could reflect that they are not binding at lower assignment levels, or that it is more costly to transfer patients under greater time pressure. Further, components such as peer image that would be inflated at λ^{-1}, which also increases with time (Figure 10). However, an implicit assumption in these calculations is that physicians do not change their transfer policy with greater workloads at EOS. I confirm this in sample, and my counterfactuals only extend work completion by at most one hour, but this would be increasingly tenuous under greater workloads.

[38]Other definitions have described presenteeism as showing up to work while ill.

[39]See, for example, definitions in the *McGraw-Hill Dictionary of American Idioms* and the *American Heritage Dictionary of Phrasal Verbs.*

the negative connotation of these terms, I argue that informational frictions imply some slacking off – potentially a significant amount – in second-best optimal assignment, which may explain why the practice is not only prevalent but also tolerated. That is, allowing workers to go home at an earlier time *ex ante* while holding work constant, or assigning more work to them while they are present, could worsen distortions further.

Social and Behavioral Mechanisms of Distortion. It is standard to assume that workers care about their own income and leisure more than the productive consequences of their workplace actions. Therefore, a natural interpretation of distortions near EOS is that they arise from strategic behavior, or moral hazard. However, other mechanisms could lead to the same welfare-reducing distortions and would have equivalent implications on how availability should be scheduled and worked assigned. For example, social norms may be that workers should not stay too long after EOS (e.g., doing so would signal incompetence), so that moving EOS too early without tapering work generates the same inefficient time pressure.[40] Workers may take schedules as a contractual "reference point" to be adhered to (Hart and Moore, 2008), and there may even be accepted routines (e.g., sign-out rounds) that reinforce this sense. Finally, although rational and forward-looking workers can always stay longer past EOS (and plan their extra-work activities accordingly), workers in practice may underestimate the time it takes to complete work, paying too much attention to when EOS is officially set.

Price and Budget Policies. It is reasonable to ask whether this inefficiency can be mitigated by price or budget policies. For example, in the conceptual framework, a wage in the form of overtime pay at $(1-\lambda)\,\tilde{c}'_\tau$ per hour would exactly cancel out any distortionary incentive near EOS. Similarly, one might speculate whether a global budget on spending for each physician might restrain the incentive to overutilize formal resources and admit patients near EOS. Under certainty and perfect information, prices (e.g., wages, or costs imposed on physicians for utilization) and quantities (e.g., physician hours) are equivalent. However, under uncertainty, control via quantities can be a superior when benefits are more concave than costs are convex, which characterizes production within most organizations at least in the short term (Weitzman,

[40] A related distortion caused by social incentives is that ED physicians care more about their peers than inpatient doctors they would be admitting their patients to. This is one (distortionary) source of client-worker specificity.

1974). Under asymmetric information and moral hazard, price or budget mechanisms are even worse. For example, under a global budget, physicians have greater incentive to cherry-pick healthier patients to stay within budget. If physicians already have the right incentives to care for patients outside the scheduling distortion (i.e., $V(\theta, d)$ and $c(\mathbf{z})$ are appropriately weighted), then a global budget could distort care uniformly toward underprovision.

9 Conclusion

I examine ED physicians working in shifts and find evidence consistent with behavioral distortions due to scheduled work: On an extensive margin, physicians are less likely to accept new patients near EOS. On an intensive margin, physicians complete their work earlier as end of shift (EOS) approaches. As the input of time becomes more costly, physicians modify the mix of inputs in patient care, and as they produce less information for discharge decisions, they are more likely to admit patients. This increases per-patient hospital costs by 21% in the last hour prior to EOS.

The EOS phenomenon documented in this paper reflects a definitional issue of scheduled work: Although scheduled availability begins and ends at set times, the true nature of work usually blurs across these constructed boundaries. Further, *ex post* worker-task specificity is often substantial in work that is information-rich. I show a tradeoff between extensive and intensive margins of distortion. In fact, observed patterns of "presenteeism" or "slacking off" may indeed be approximately second-best optimal. Key to this result is that physicians are willing to spend increasingly large amounts of hospital dollars for each hour of their leisure time, a finding that sheds light on the tradeoff between intrinsic and extrinsic motivations. This is relevant for a wide set of policy levers that act via assignment. Attempting to prevent workers from sitting idly could be quite costly when used at the wrong time in scheduled work.

References

Aghion, P. and J. Tirole, "Formal and real authority in organizations," *The Journal of Political Economy*, 1997, *105* (1), 1–29.

Akerlof, George A. and Rachel E. Kranton, "Identity and the Economics of Organizations," *The Journal of Economic Perspectives*, January 2005, *19* (1), 9–32.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber, "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, February 2005, *113* (1), 151–184.

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.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul, "Social preferences and the response to incentives: Evidence from personnel data," *The Quarterly Journal of Economics*, 2005, *120* (3), 917–962.

_ , _ , and _ , "Social connections and incentives in the workplace: Evidence from personnel data," *Econometrica*, 2009, *77* (4), 1047–1094.

Beers, Thomas, "Flexible schedules and shift work: Replacing the 9-to-5 workday?," Technical Report, Bureau of Labor Statistics June 2000.

Benabou, Roland and Jean Tirole, "Intrinsic and Extrinsic Motivation," *The Review of Economic Studies*, July 2003, *70* (3), 489–520.

Besley, Timothy and Maitreesh Ghatak, "Competition and Incentives with Motivated Agents," *American Economic Review*, 2005, *95* (3), 616–636.

Bloom, N., J. Liang, J. Roberts, and Z. J. Ying, "Does Working from Home Work? Evidence from a Chinese Experiment," *The Quarterly Journal of Economics*, November 2014.

Brachet, Tanguy, Guy David, and Andrea M Drechsler, "The Effect of Shift Structure on Performance," *American Economic Journal: Applied Economics*, April 2012, *4* (2), 219–246.

Briscoe, Forrest, "Temporal Flexibility and Careers: The Role of Large-Scale Organizations for Physicians," *Industrial and Labor Relations Review*, 2006, *60*, 88.

_ , "From Iron Cage to Iron Shield? How Bureaucracy Enables Temporal Flexibility for Professional Service Workers," *Organization Science*, April 2007, *18* (2), 297–314.

Casalino, Lawrence P., Kelly J. Devers, Timothy K. Lake, Marie Reed, and Jeffrey J. Stoddard, "Benefits of and barriers to large medical group practice in the united states," *Archives of Internal Medicine*, September 2003, *163* (16), 1958–1964.

Chan, David, "Teamwork and Moral Hazard: Evidence from the Emergency Department," *Journal of Political Economy*, 2015, *Forthcoming.*

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff, "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, 2014, *104* (9), 2593–2632.

Coviello, Decio, Andrea Ichino, and Nicola Persico, "Time Allocation and Task Juggling," *The American Economic Review*, 2014, *104* (2), 609–623.

Dewatripont, Mathias, Ian Jewitt, and Jean Tirole, "The Economics of Career Concerns, Part II: Application to Missions and Accountability of Government Agencies," *The Review of Economic Studies*, January 1999, *66* (1), 199–217.

Elixhauser, Anne, Claudia Steiner, D. Robert Harris, and Rosanna M. Coffey, "Comorbidity Measures for Use with Administrative Data," *Medical Care*, January 1998, *36* (1), 8–27.

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

Forster, Alan J., Ian Stiell, George Wells, Alexander J. Lee, and Carl Van Walraven, "The Effect of Hospital Occupancy on Emergency Department Length of Stay and Patient Disposition," *Academic Emergency Medicine*, 2003, *10* (2), 127–133.

Garicano, Luis, "Hierarchies and the Organization of Knowledge in Production," *Journal of Political Economy*, October 2000, *108* (5), 874–904.

Goldin, Claudia, "A Grand Gender Convergence: Its Last Chapter," *American Economic Review*, April 2014, *104* (4), 1091–1119.

Green, Linda, "A Multiple Dispatch Queueing Model of Police Patrol Operations," *Management Science*, June 1984, *30* (6), 653–664.

Green, Linda V., "Capacity Planning and Management in Hospitals," in Margaret L. Brandeau, FranÃ§ois Sainfort, and William P. Pierskalla, eds., *Operations Research and Health Care*, number 70. In 'International Series in Operations Research & Management Science.', Springer US, January 2004, pp. 15–41.

Hart, Oliver and Bengt Holmstrom, "The theory of contracts," in Truman F. Bewley, ed., *Advances in Economic Theory*, Cambridge, UK: Cambridge University Press, 1987, pp. 71–156.

_ **and John Moore**, "Contracts as Reference Points," *The Quarterly Journal of Economics*, February 2008, *123* (1), 1–48.

He, Biyu, Franklin Dexter, Alex Macario, and Stefanos Zenios, "The Timing of Staffing Decisions in Hospital Operating Rooms: Incorporating Workload Heterogeneity into the Newsvendor Problem," *Manufacturing and Service Operations Management*, January 2012, *14* (1), 99–114.

Ichniowski, Casey, Thomas A. Kochan, David Levine, Craig Olson, and George Strauss, "What Works at Work: Overview and Assessment," *Industrial Relations: A Journal of Economy and Society*, July 1996, *35* (3), 299–333.

Jacob, Brian A., Lars Lefgren, and David P. Sims, "The Persistence of Teacher-Induced Learning," *Journal of Human Resources*, September 2010, *45* (4), 915–943.

Kc, Diwas S. and Christian Terwiesch, "Impact of workload on service time and patient safety: an econometric analysis of hospital operations," *Management Science*, September 2009, *55* (9), 1486–1498.

Kolstad, Jonathan T., "Information and Quality When Motivation Is Intrinsic: Evidence from Surgeon Report Cards," *American Economic Review*, 2013, *103* (7), 2875–2910.

Lambert, Diane, "Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing," *Technometrics*, February 1992, *34* (1), 1–14.

Larkin, Ian, "The cost of high-powered incentives: Employee gaming in enterprise software sales," *Journal of Labor Economics*, April 2014, *32* (2), 199–227.

Lerman, Benjamin and Michael S. Kobernick, "Return Visits to the Emergency Department," *The Journal of Emergency Medicine*, September 1987, *5* (5), 359–362.

Liebman, Jeffrey B. and Neale Mahoney, "Do Expiring Budgets Lead to Wasteful Year-End Spending? Evidence from Federal Procurement," Working Paper 19481, National Bureau of Economic Research September 2013.

Maher, Kris, "Wal-mart seeks flexibility in worker shifts," *The Wall Street Journal*, January 2007.

Marschak, Jacob and Roy Radner, *Economic Theory of Teams*, New Haven, CT: Yale University Press, 1972.

Mas, Alexandre and Enrico Moretti, "Peers at Work," *The American Economic Review*, 2009, *99* (1), 112–145.

Messerli, Franz H., Adrian W. Messerli, and Thomas F. LÃŒscher, "Eisenhower's Billion-Dollar Heart Attack – 50 Years Later," *New England Journal of Medicine*, September 2005, *353* (12), 1205–1207.

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

Oyer, Paul, "Fiscal year ends and nonlinear incentive contracts: The effect on business seasonality," *The Quarterly Journal of Economics*, 1998, *113* (1), 149–185.

Papke, Leslie E. and Jeffrey M. Wooldridge, "Econometric Methods for Fractional Response Variables with an Application to 401(K) Plan Participation Rates," *Journal of Applied Econometrics*, 1996, *11* (6), 619–632.

Perdikaki, Olga, Saravanan Kesavan, and Jayashankar M. Swaminathan, "Effect of Traffic on Sales and Conversion Rates of Retail Stores," *Manufacturing and Service Operations Management*, January 2012, *14* (1), 145–162.

Prendergast, Canice, "The Motivation and Bias of Bureaucrats," *The American Economic Review*, March 2007, *97* (1), 180–196.

Presser, Harriet B., *Working in a 24/7 Economy: Challenges for American Families.*, Russell Sage Foundation, 2003.

Radner, Roy, "The Organization of Decentralized Information Processing," *Econometrica*, 1993, *61* (5), 1109–46.

Schuur, Jeremiah D. and Arjun K. Venkatesh, "The Growing Role of Emergency Departments in Hospital Admissions," *The New England Journal of Medicine*, 2012, *367*, 391–393.

Shapiro, Carl and Joseph E. Stiglitz, "Equilibrium Unemployment as a Worker Discipline Device," *The American Economic Review*, June 1984, *74* (3), 433–444.

Shetty, Kanaka D. and Jayanta Bhattacharya, "Changes in hospital mortality associated with residency work-hour regulations," *Annals of Internal Medicine*, July 2007, *147* (2), 73–80.

Simon, H.A., *Administrative Behavior*, New York: Macmillan, 1947.

Simpson, Ruth, "Presenteeism, Power and Organizational Change: Long Hours as a Career Barrier and the Impact on the Working Lives of Women Managers," *British Journal of Management*, September 1998, *9*, 37–50.

Tanabe, Paula, Rick Gimbel, Paul R. Yarnold, Demetrios N. Kyriacou, and James G. Adams, "Reliability and Validity of Scores on the Emergency Severity Index Version 3," *Academic Emergency Medicine*, 2004, *11* (1), 59–65.

Tirole, J., "Hierarchies and bureaucracies: On the role of collusion in organizations," *Journal of Law, Economics and Organization*, 1986, *2*, 181–214.

Volpp, K. G. and A. K. Rosen, "Mortality among hospitalized Medicare beneficiaries in the first two years following ACGME resident duty hour reform," *The Journal of the American Medical Association*, September 2007, *298* (9), 975–983.

Weitzman, Martin L., "Prices vs. Quantities," *The Review of Economic Studies*, October 1974, *41* (4), 477–491.

Table 1: End of Shift Effect on Log Length of Stay

	(1)	(2)	(3)	(4)	(5)
	Log length of stay				
Hour prior to EOS					
Last hour	-0.607***	-0.547***	-0.529***	-0.716***	-0.587***
	(0.028)	(0.025)	(0.025)	(0.039)	(0.050)
Second hour	-0.316***	-0.282***	-0.330***	-0.461***	-0.287***
	(0.008)	(0.008)	(0.008)	(0.012)	(0.026)
Third hour	-0.139***	-0.129***	-0.161***	-0.260***	-0.123***
	(0.005)	(0.005)	(0.006)	(0.009)	(0.022)
Fourth hour	-0.112***	-0.092***	-0.111***	-0.173***	-0.091***
	(0.005)	(0.004)	(0.005)	(0.008)	(0.018)
Fifth hour	-0.070***	-0.055***	-0.078***	-0.120***	-0.023
	(0.004)	(0.004)	(0.005)	(0.007)	(0.015)
Sixth hour	-0.065***	-0.048***	-0.057***	-0.090***	-0.010
	(0.004)	(0.004)	(0.005)	(0.007)	(0.012)
Patient characteristics	N	Y	Y	Y	Y
Time and pod dummies	N	N	Y	Y	Y
Physician-resident-nurse identities	N	N	N	Y	Y
Time relative to shift beginning	N	N	N	N	Y
Number of observations	371,107	371,107	371,107	371,107	371,107
Adjusted R-squared	0.008	0.189	0.211	0.400	0.410
Sample mean log length of stay (log hours)	1.050	1.050	1.050	1.050	1.050

Note: This table reports coefficient estimates and standard errors in parentheses for versions of Equation (2) regressing log length of stay, with increasing controls, for arrival at each hour prior to end of shift (EOS), where arrival greater than six hours is the reference period. Patient characteristics include demographics, emergency severity index (ESI), time spent in triage, and rich indicators for clinical diagnoses (e.g., Elixhauser indices). Time dummies include indicators for hour of day, day of week, and month-year interactions. * denotes significance at 10% level, ** denotes significance at 5% level, and *** denotes significance at 1% level.

Table 2: End of Shift Effect on Other Outcomes

	(1)	(2)	(3)	(4)	(5)	(6)
	Order count	Inpatient admission	Log total cost	30-day mortality	14-day bounce-back	Workload-adjusted LOS
Hour prior to EOS						
Last hour	1.411**	0.057**	0.208**	-0.003	-0.028	-0.144***
	(0.562)	(0.024)	(0.080)	(0.008)	(0.018)	(0.051)
Second hour	-0.093	0.000	0.027	-0.001	-0.011	0.015
	(0.302)	(0.013)	(0.043)	(0.004)	(0.010)	(0.027)
Third hour	-0.003	0.002	0.009	-0.005	-0.005	0.090***
	(0.249)	(0.011)	(0.036)	(0.004)	(0.008)	(0.022)
Fourth hour	0.167	0.004	0.029	-0.001	-0.002	0.036*
	(0.207)	(0.009)	(0.030)	(0.003)	(0.007)	(0.018)
Fifth hour	0.239	-0.004	0.034	-0.002	0.001	0.037**
	(0.171)	(0.007)	(0.024)	(0.003)	(0.005)	(0.015)
Sixth hour	0.192	-0.007	-0.006	0.001	0.001	0.001
	(0.137)	(0.006)	(0.019)	(0.002)	(0.004)	(0.012)
Number of observations	371,421	371,421	366,219	371,421	371,421	371, 148
Adjusted R-squared	0.531	0.459	0.472	0.295	-0.044	0.476
Sample mean outcome	13.518	0.269	6.750	0.018	0.060	-0.904

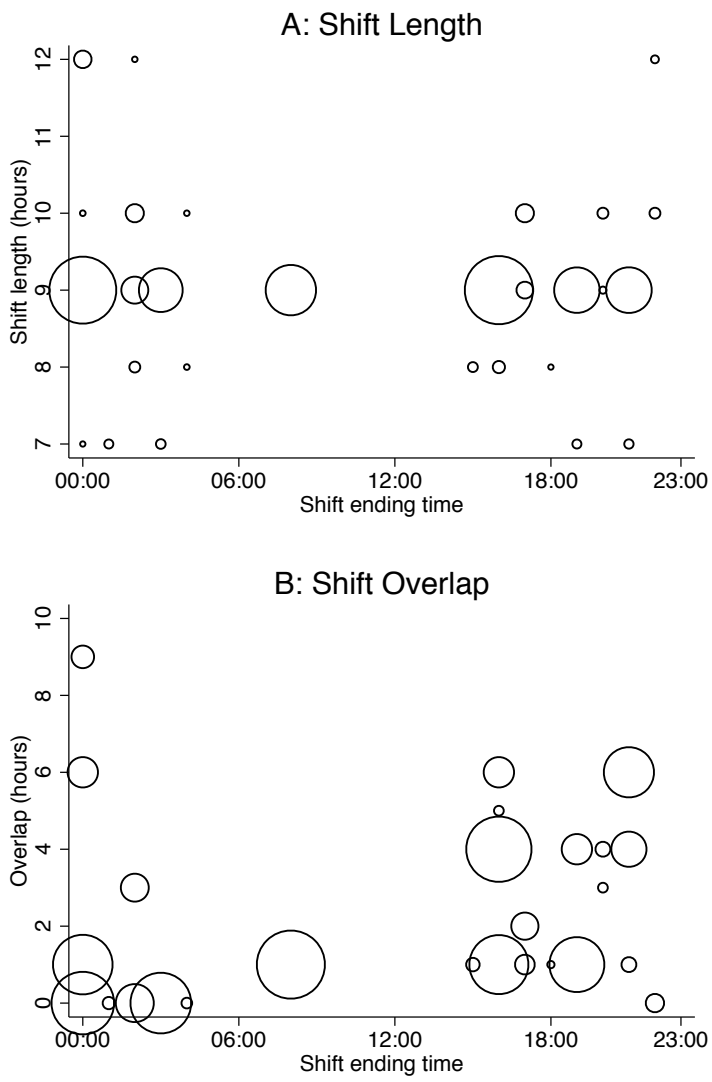
Note: This table reports coefficient estimates and standard errors in parentheses for Equation (2) with a full set of controls regressing other outcome variables, for arrival at each hour prior to end of shift (EOS), where arrival greater than six hours is the reference period. Workload-adjusted length of stay (LOS) is calculated by Equation (5). Controls are as described for Table 1. * denotes significance at 10% level, ** denotes significance at 5% level, and *** denotes significance at 1% level.

Table 3: Effect on Workload-adjusted Length of Stay by Shift Overlap

	(1)		(2)	(3)
	$\overline{o} \leq 1$	$\overline{o} \geq 2$	$\overline{o} \leq 1$	$\overline{o} \geq 2$
Hour prior to EOS				
Last hour	-0.167**	-0.034	-0.229***	-0.003
	(0.068)	(0.110)	(0.069)	(0.143)
Second hour	0.015	0.126*	0.014	0.140
	(0.038)	(0.067)	(0.039)	(0.085)
Third hour	0.05	0.137**	0.037	0.099
	(0.031)	(0.056)	(0.033)	(0.069)
Fourth hour	0.007	0.088*	-0.002	0.056
	(0.025)	(0.051)	(0.026)	(0.058)
Fifth hour	0.013	0.085*	0.009	0.052
	(0.021)	(0.044)	(0.022)	(0.047)
Sixth hour	-0.017	0.044	-0.022	0.029
	(0.015)	(0.038)	(0.016)	(0.031)
Control for time relative to shift beginning	Y		Y	Y
Patient, provider, and other time controls	Y		Y	Y
Sample	Full, actual		$\overline{o} \leq 1$, actual	$\overline{o} \geq 2$, actual
Number of observations	333,233		231,576	101,657
Adjusted R-squared	0.456		0.491	0.502
Sample mean outcome	-0.926		-0.987	-0.789

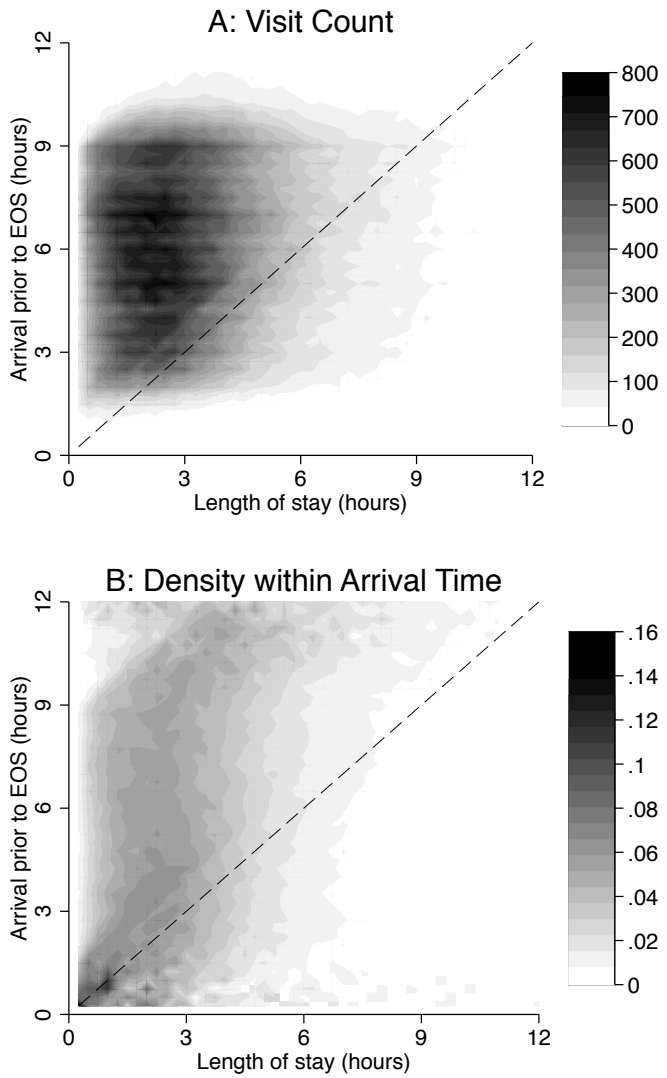
Note: This table reports coefficient estimates and standard errors in parentheses for EOS effects on workload-adjusted length of stay, for arrival at each hour prior to end of shift (EOS), where arrival greater than six hours is the reference period. Model (1) is estimated by Equation (4), while models (2) and (3) are estimated separately by Equation (2) on subsamples of the data according to $\overline{o}$. All three models are estimted with a full set of controls, as described for Table 1. Workload-adjusted length of stay is calculated by Equation (5). * denotes significance at 10% level, ** denotes significance at 5% level, and *** denotes significance at 1% level. This table is continued by Table A-5.3.

Figure 1: Shift Variation



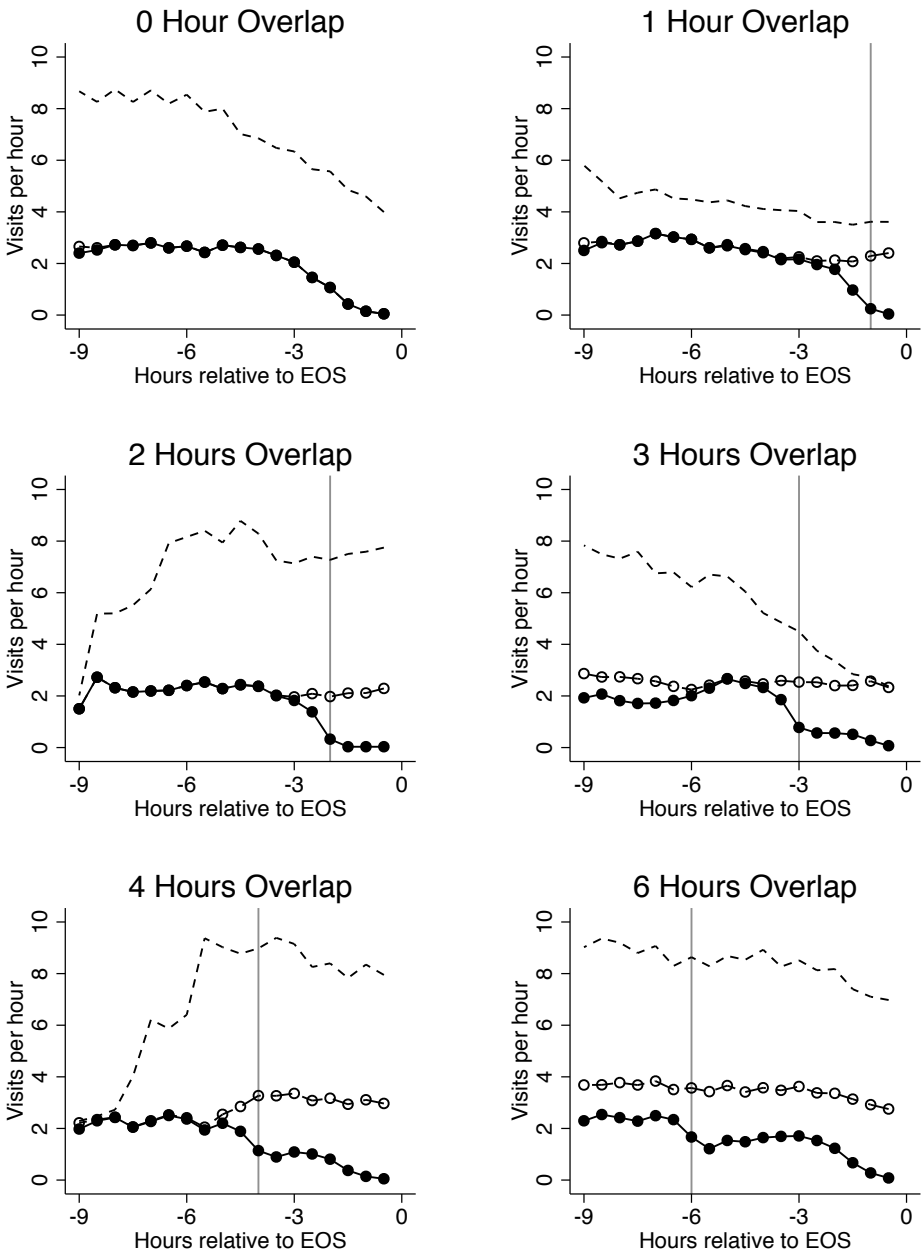
Note: This figure illustrates the variation in observations across shift types. Panel A plots shifts by shift ending time and shift length. Panel B plots shifts by shift ending time and the length of overlapping transition at the end of shift.

Figure 2: Density of Visits on Arrival Time and Length of Stay



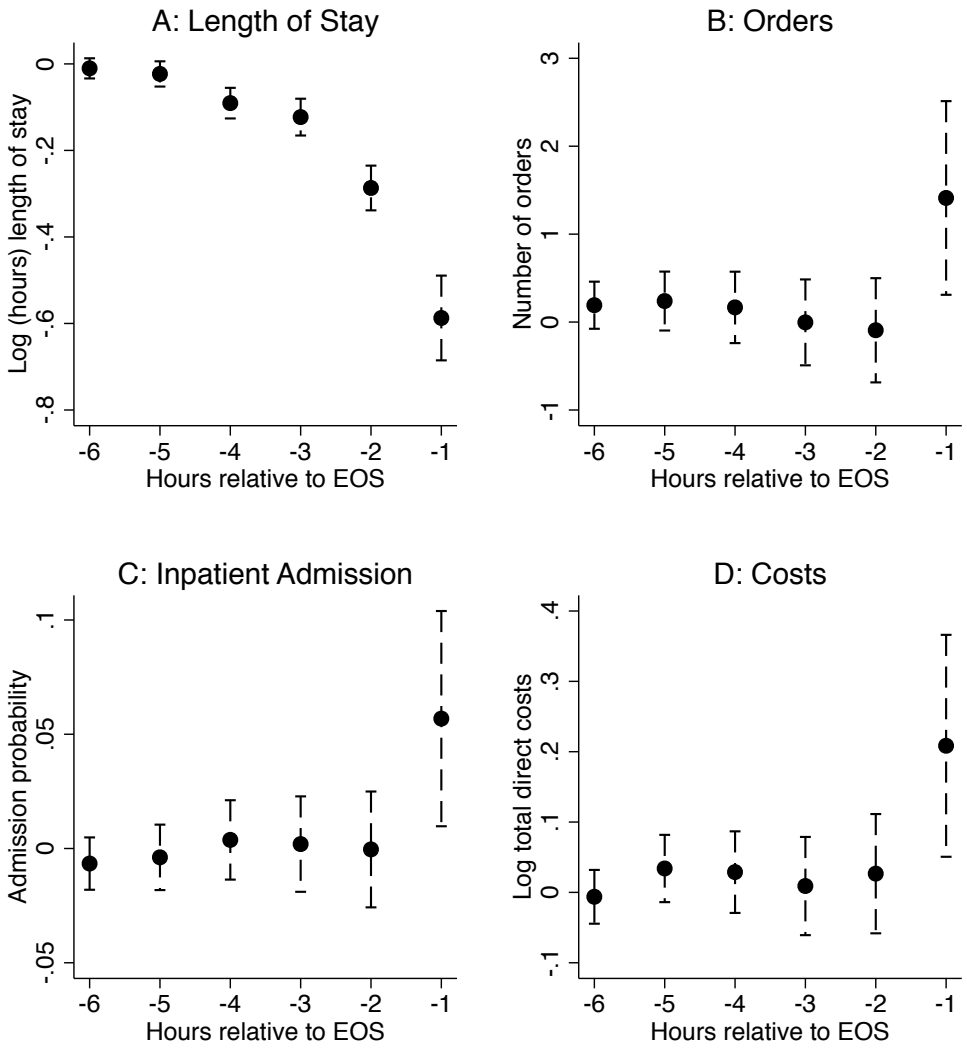
Note: This figure plots the distribution of visits over arrival times relative to EOS and length of stay. Panel A plots visit counts within fifteen-minute intervals of arrival time and length of stay. Panel B plots the density of visits, conditional on arrival time.

Figure 3: Flow of Patient Visits over Time



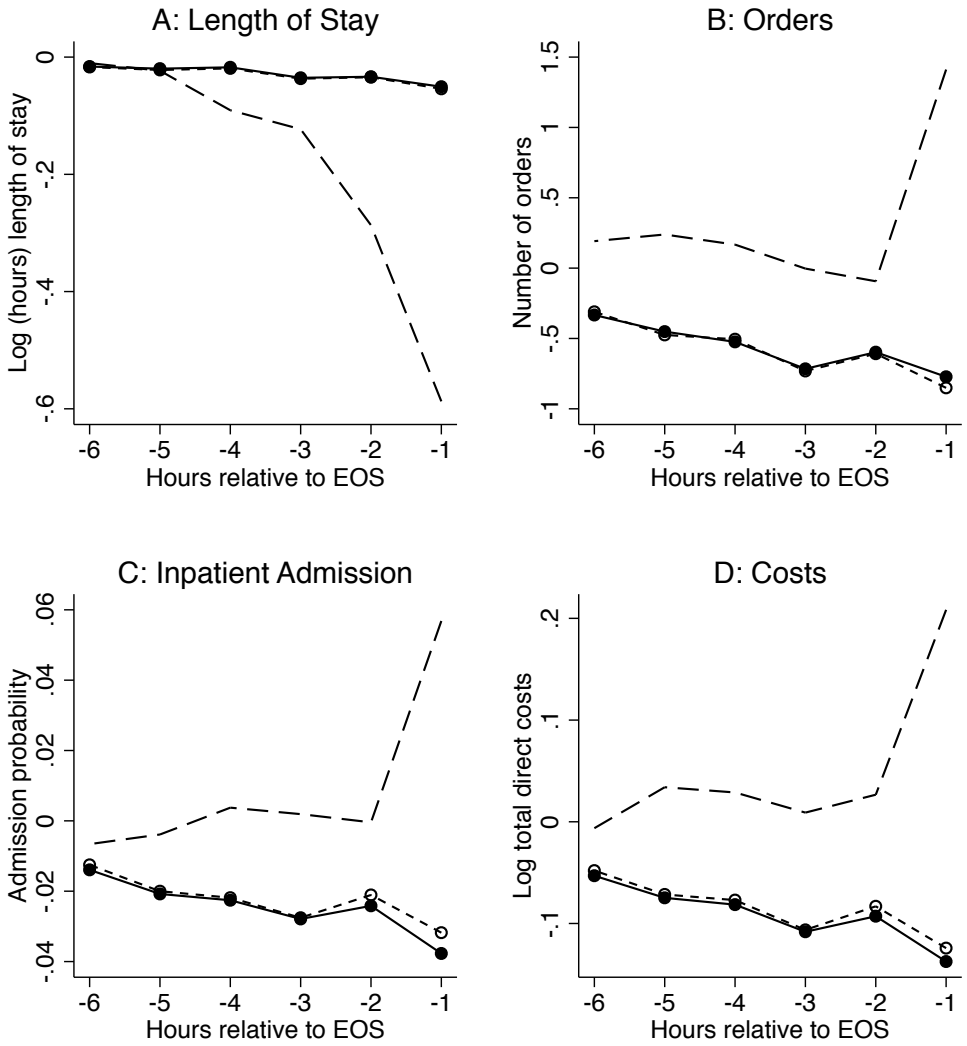
Note: This figure shows unadjusted average hourly rates of patient visits for each 30-minute interval relative to end of shift (EOS). Each panel shows results for shifts with a given EOS overlap time. Patient visits for the index physician are shown in closed circles; patient visits for the location are shown in open circles; and patient visits for the entire ED are shown with a dashed line with no markers. Subsequent shift starting times are marked with a vertical line.

Figure 4: End of Shift Effects



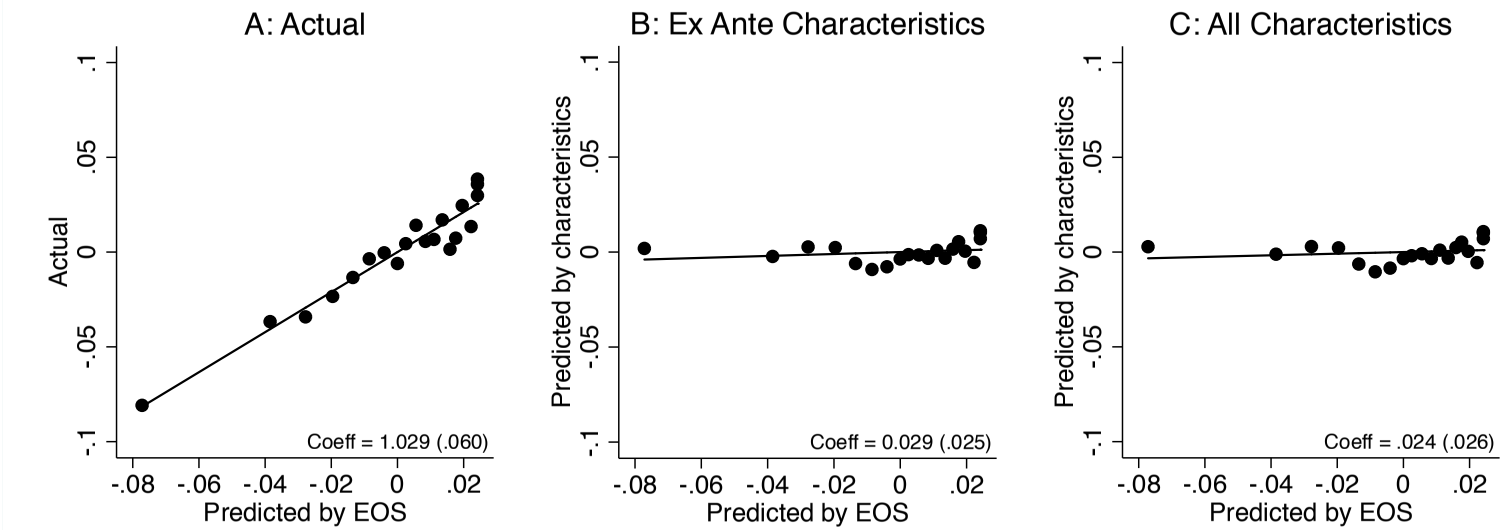
Note: This figure plots average effects for each hour prior to end of shift (EOS) on length of stay (Panel A), orders (Panel B), inpatient admissions (Panel C), and costs (Panel D). Each outcome is estimated separately using Equation (5). The reference category is any time greater than six hours prior to EOS. Bracketed dashed lines represent 95% confidence intervals for each estimate.

Figure 5: Patient Selection on Observables Relative to End of Shift



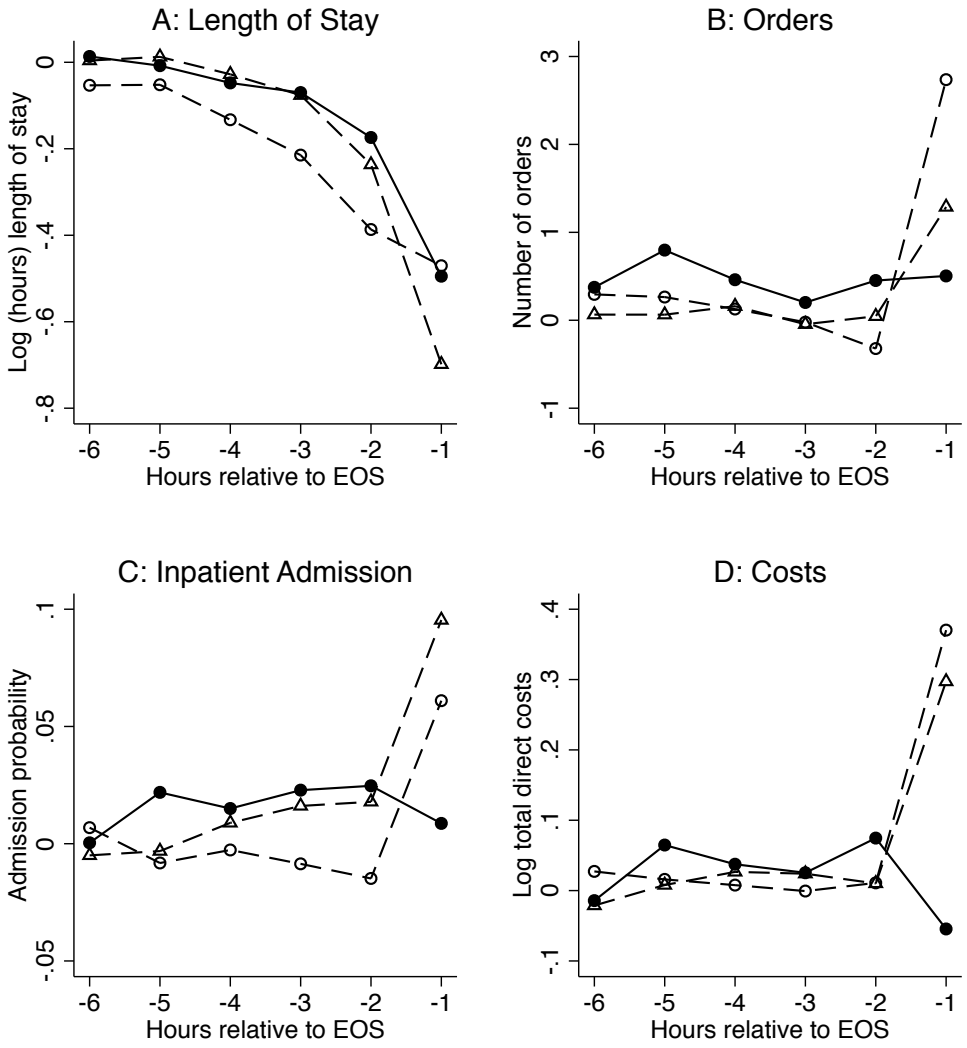
Note: This figure shows selection on observables for each hour prior to end of shift (EOS) on length of stay (Panel A), orders (Panel B), inpatient admissions (Panel C), and costs (Panel D). Each outcome is predicted based on patient characteristics observable prior to acceptance (age, sex, ESI) (closed circles) and on the full set of characteristics usually unobservable until after patient acceptance (e.g., 29 Elixhauser indices, race, language) (short-dashed line, open circles). Coefficients are estimated for predicted outcome using Equation (A-3.2). For reference, adjusted effects on actual outcomes from Figure 4 are shown with the dashed line. The reference category is any time greater than six hours prior to EOS.

Figure 6: Selection Within and Across Hours



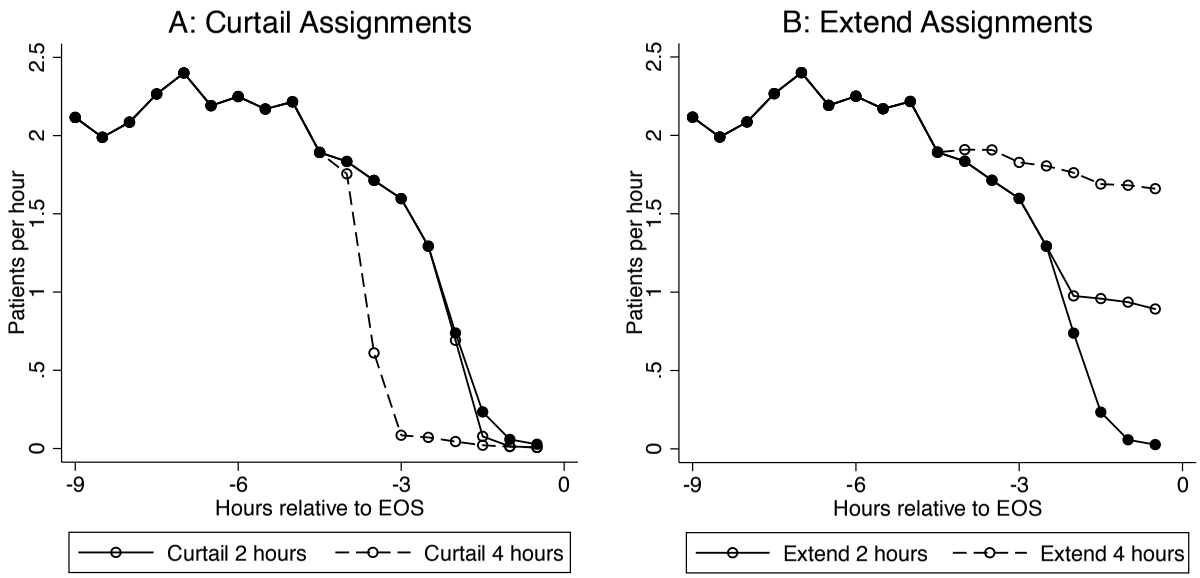
Note: This figure shows binned scatterplots of actual (residualized) log length of stay (Panel A), log length of stay predicted on "ex ante" characteristics usually observable to physicians prior to acceptance (Panel B), and log length of stay predicted on all characteristics including those usually observable only after acceptance (Panel C). Predicted and actual log lengths of stay are all averaged within hour cell and weighted by visit. The core data for the x-axis on all three panels is the log length of stay predicted by the times to EOS, defined by Equation (A-3.8) as Q_t. $Q_t$is calculated as follows: First, coefficients on time relative to EOS are calculated from (2) using leave-shift-out sampling. Next, these coefficients are averaged across shifts in process at hour t, weighted by visits.To calculate residualized actual log length of stay (Panel A), I subtract expected log length of stay based on all covariates listed in the note for Table 1, except for time to EOS, using only variation within time to EOS. To calculate predicted log length of stay by patient characteristics (Panels B and C), I residualize the characteristics by time categories and use within-EOS-time variation to predict log length of stay. Patient characteristics and time categories are described in the notes for Figure 5 and Table 1, respectively. To construct each of the binned scatterplots, I demean values on the x- and y-axis, separate the data into 20 equal-sized groups (by patient visits) ordered by A_t, then plot the mean value within each bin. Solid lines show the best linear fit by OLS on the underlying microdata, clustered by hour (coefficients and standard errors are given as notes in each panel, also given in Table A-3.2). The same data in this figure is also presented in Figure A-3.9. Details are given in Appendix A-3.4.

Figure 7: End of Shift Effects by Shift Overlap



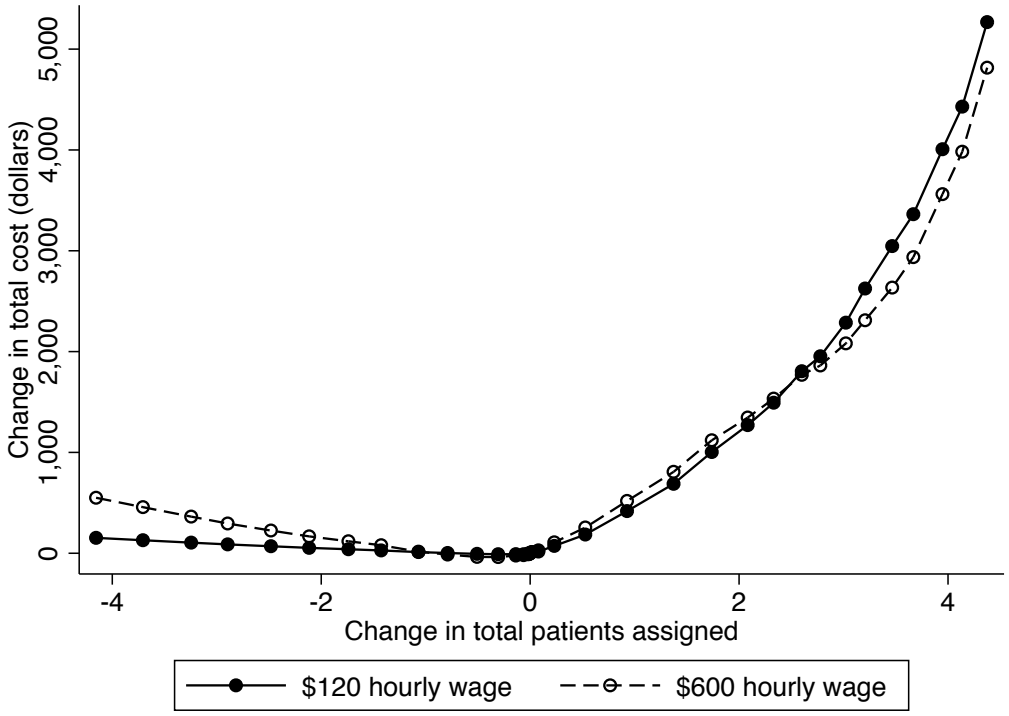
Note: This figure shows heterogeneous end of shift (EOS) effects by EOS overlap times on length of stay (Panel A), orders (Panel B), inpatient admissions (Panel C), and costs (Panel D). Each outcome is estimated separately using Equation (4). Estimates for terminal shifts ($\overline{o} = 0$) are shown in open triangles; estimates for minimally transitioned shifts ($\overline{o} = 1$) are shown in open circles; and estimates for substantially transitioned shifts ($\overline{o} \geq 2$) are shown in closed circles.

Figure 8: Example Counterfactual Assignment Regimes



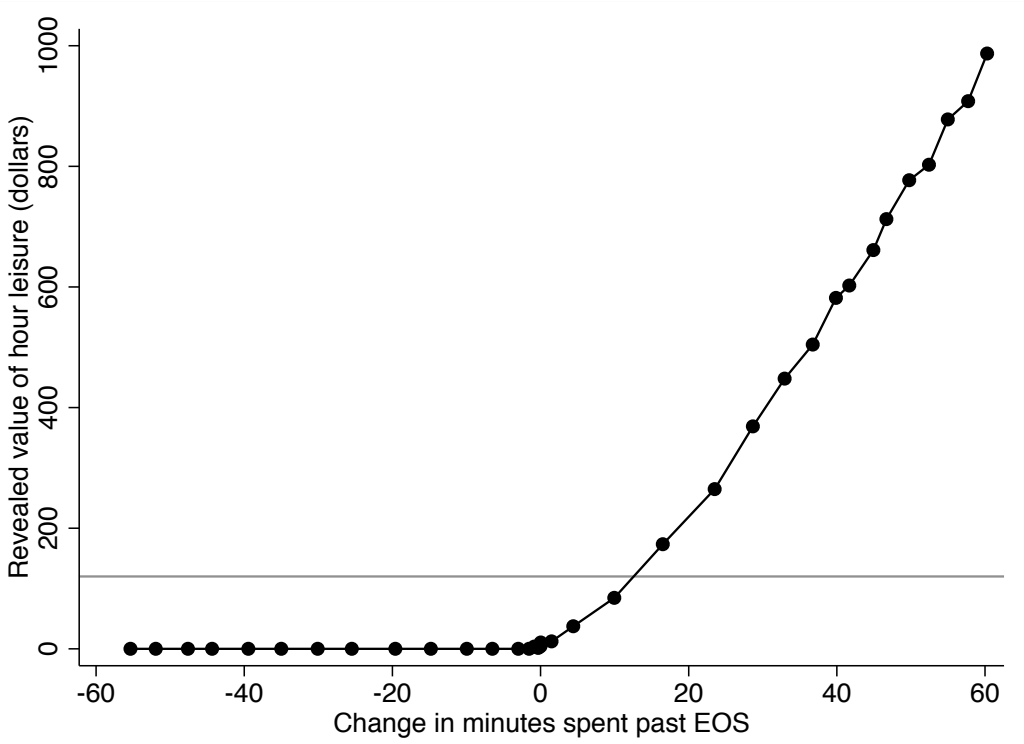
Note: This figure shows example counterfactual assignment regimes, parameterized as hours Δ curtailing or extending assignment, as specified by Equation (A-4.17). Panel A shows counterfactual regimes in which assignment is curtailed earlier than actual assignment patterns. Panel B shows counterfactual regimes in which assignment is extended beyond actual assignment patterns. Two and four hours denote the times at which curtailment or extension begins.

Figure 9: Change in Total Cost per Shift over Counterfactual Regimes



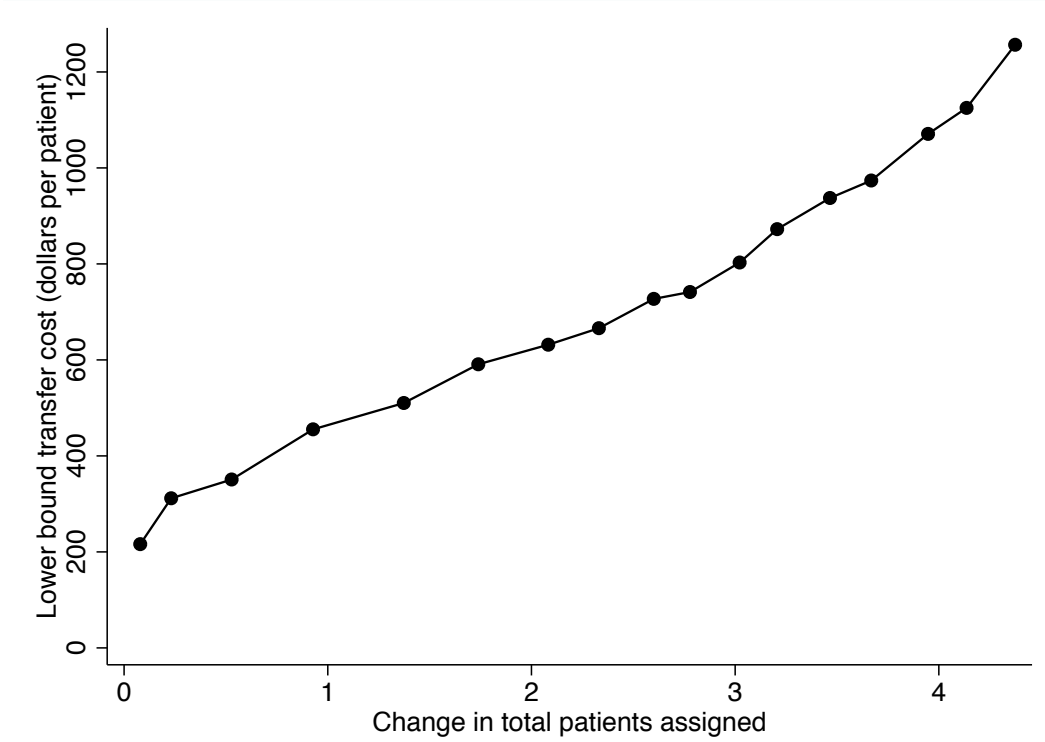
Note: This figure plots changes in total cost during a shift averaged over 100 simulations for each counterfactual assignment regime, as described in Section 7 and Appendix A-4. Assignment regimes may either curtail or extend assignment, as illustrated in Figure 8. The x-axis is the change in total patients assigned during a shift as a result of an assignment regime ("0" represents the actual assignment regime, which by definition has a value of 0 on the y-axis). Daily costs include both physician-time, patient-time, and hospital-resource costs. Changes in daily cost are plotted under the base-case assumption of a \$120 hourly wage (solid circles) and an extreme case of a \$600 hourly wage (hollow circles).

Figure 10: Value of Leisure in Dollars of Patient Care



Note: This figure plots the imputed value of leisure, revealed by increases in resource-utilization costs that shorten the time for completion of work under simulated counterfactual assignment policies. For each counterfactual assignment policy, data are simulated using the actual discharge policy and then again for a counterfactual discharge policy that is insensitive to time relative to EOS. The difference between these two discharge policies yields a tradeoff of resource-utilization costs for shortened work completion time. The ratio between these two represents the value of leisure, denominated in hospital-resource dollars, and is plotted on the y-axis. The x-axis is the change work completion time using the actual discharge policy. The horizontal line drawn at \$120 is the effective hourly wage for an hour of a physician's (scheduled) time. Details are described in Appendix A-4.

Figure 11: Lower Bound of Transfer Cost



Note: This figure plots a lower bound of the cost of transferring patients to another physician at EOS. This lower bound is given by the logic that increases in patient-costs could be eliminated if transfer costs were zero and reduced if transfer costs were lower (by transferring the additional patients and incurring the costs). For each counterfactual assignment policy that increases assigned patients over the actual policy, I calculate this lower bound by dividing increases in resource-utilization costs with increases in assigned patients. The x-axis shows the increase in assigned patients under a counterfactual regime; the y-axis shows the per-patient lower bound transfer cost in dollars. Given a base per-patient cost of \$894, this lower bound ranges from 25% to 140%. Details are described in Appendix A-4.

Appendix

A-1 Effects Relative to Shift Beginning

The literature on shift work has almost exclusively focused on cumulative health effects and fatigue (e.g., Brachet et al., 2012; Shetty and Bhattacharya, 2007; Volpp and Rosen, 2007), while I explore the possibility of strategic behavior in this paper. Unlike shifts of 36 hours in the residency work-hours debate, significant fatigue is less likely near the end of a shift of nine hours, the modal shift length in this setting. Nonetheless, I specifically address this issue by exploiting variation in shift length to control for effects, such as fatigue, correlated with time since the beginning of shift. I assume that, conditional on time since beginning of shift, fatigue is independent of time to EOS.

In the full model of Equation (2), I show robust EOS effects controlling for time since the beginning of shift. The effect attributable to time since shift beginning is minor compared to the overall effect for length of stay. Here I illustrate the robustness of EOS effects more directly by simply showing the effect on length of stay for each hour prior to EOS separately for three categories of shift lengths. I study shifts that are nine hours in length, as well as shifts that are seven or eight hours in lengths and shifts that are ten hours in length. Figure A-1.1 plots coefficients α_m from Equation (2) estimated separately for each shift-length category. Panel A plots coefficients according to time relative to EOS and shows coefficients largely similar across shift lengths and within hour prior to EOS. Panel B arranges the coefficients according to time from shift beginning, illustrating the corollary that the EOS effect is largely independent of the time since beginning the shift.

A-2 Time Components of Length of Stay

In Section 5, length of stay decreases while formal utilization increases near EOS. This suggests that formal utilization is a net substitute for time in patient care. In this appendix, I further examine this hypothesis by a closer look at the time components of length of stay. In practice, time is not neatly divided into pure substitute or complement components with formal utilization (call these components τ_1 and τ_2, respectively), but some intuitive distinctions can be made: Time before the first formal order likely belongs to τ_1 (e.g., time spent interviewing the patient or performing serial abdominal examination as opposed to CT scan). Time after the last formal order likely belongs to τ_2, reflecting time needed to follow up on utilization (e.g., waiting for CT scan report). Although time in between the first and last orders could belong to either τ_1 or τ_2, the spacing of these orders often reflects clinical monitoring and reasoning more closely related to τ_1.

Measuring length of stay in three component shares – time between pod arrival and first order, time between first and last (non-discharge) orders, and time between last and discharge

orders – I estimate a fractional logit model (Papke and Wooldridge, 1996) using similar regressors as in Equation (2). Figure A-2.1 presents results of marginal effects relative to EOS. Panel A scales time shares by the median predicted length of stay in each hour prior to EOS according (2); Panel B simply plots the unscaled proportional shares. These proportions remain relatively unchanged except for the last hour prior to EOS, when the proportions for time prior to first order and inter-order time both decrease. These results suggest relative reductions in τ_1, particularly in the last hour prior to EOS, and are consistent with the increase in formal utilization (net substitution) in the last hour shown in Table 2 and Figure 4.

A-3 Selection of Patient Types

A-3.1 Summary Statistics of Observables

I first present plots of summary statistics of observable characteristics of accepted patients arriving in each 30-minute interval relative to EOS. Figures A-3.1 to A-3.5 present mean age, mean ESI, proportion white race, proportion black race, and proportion Spanish-speaking, respectively. Selection is towards healthier patients (or disadvantaged patients, who tend to visit the ED for less serious reasons) for all of these measures but is small.

In Figures A-3.1 to A-3.5, I also separately consider selection in shifts without overlap and in shifts without overlap. Recall that overlap is the time prior to EOS during which a physician shares new work with another physician who has begun work in the same location. Thus, shifts without overlap are "terminal" shifts in which physicians are unable to decline patients who are assigned to their managerial location, and selection must therefore occur by the triage nurse assigning different types of patients to the managerial location. Of note, selection is relatively greater in these shifts than in shifts with overlap, consistent with cultural norms against selection between physicians sharing a location.

In Figures A-3.6 and A-3.7, I plot quantiles of continuous variables age and predicted log length of stay, respectively. Predictions of log length of stay are based on cubic splines of age, an indicator for male sex, indicators of ESI, indicators for race, and indicators for language. These quantiles show persistently wide variation in the patients accepted within each 30-minute interval relative to EOS. Each of these quantiles are stable and only slightly decreasing across times relative to EOS. Thus, in addition to means, the entire distribution of these characteristics for accepted patients does not change as EOS approaches. Similarly, Figure A-3.8 shows cumulative proportions of patients by ESI. These proportions are also stable across time intervals.

A-3.2 Selection on Ex Ante Observables and Unobservables

This appendix section makes use of the fact that I observe *ex post* a richer set of patient characteristics, including diagnoses and other characteristics, than generally are unobserved by physician at patient acceptance. In this analysis, I evaluate patient selection based on two sets

of characteristics: those that are observed by a physician or triage nurse prior to acceptance, $\mathbf{X}_{it}^{prior}$, and others that include rich diagnosis codes, insurance status, race, and language that are at best incompletely observed until after patient acceptance, $\mathbf{X}_{it}^{full}$.

Separately for each set, I first generate predicted outcomes by

$$Y_{ijkpt} = \left(\mathbf{X}_{it}^{set}\right)' \beta^{set} + \varepsilon_{ijkpt}, \tag{A-3.1}$$

where $set \in \{prior, full\}$. Next, I estimate the following regression describing the relationship between the predicted outcomes for selected patients, $\hat{Y}_{ijkpt}^{set} = \hat{\beta}^{set}\mathbf{X}_{it}^{set}$, using $\hat{\beta}^{set}$ estimated from Equation (A-3.1), and the time of selection relative to EOS:

$$\begin{aligned} \hat{Y}_{ijkpt}^{set} \quad = \quad & \sum_{m=-6}^{-1} \alpha_m^{set} \mathbf{1}\left(\left\lfloor t - \overline{t}\left(j,t\right)\right\rfloor = m\right) + \sum_{m} \gamma_m \mathbf{1}\left(\left\lfloor t - \underline{t}\left(j,t\right)\right\rfloor = m\right) + \\ & \mathbf{T}_t'\eta + \zeta_p + \nu_{jk} + \varepsilon_{ijkpt}, \end{aligned} \tag{A-3.2}$$

leaving out variables in $\mathbf{X}_{it}$ as regressors. I interpret each coefficient α_m^{set} as the amount patient selection, in terms of length of stay predicted by $\mathbf{X}_{it}^{set}$. Comparing results between the two sets of patient characteristics roughly assesses the degree of selection on characteristics unobservable at the time of patient acceptance but observable to me.

Figure 5 presents estimates of selection for each set of patient characteristics and for each of the outcomes of length of stay, orders, admission, and costs. To reference magnitude, selection estimates are overlaid onto estimates for the EOS effect from Equation (2) for each respective outcome. Coefficients for selection estimated using the two sets of characteristics are remarkably similar, suggesting negligible additional selection on *ex ante* unobservables. Selection nearing EOS appears to be in the direction of healthier or less resource-intensive patients: those expected to have shorter lengths of stay, lower frequencies of admissions, and incur lower costs and fewer orders. Predicted length of stay is 5.4% lower in the last hour prior to EOS compared to seven or more hours prior to EOS, about an order of magnitude smaller than effects for actual length of stay. All predicted outcomes show a *decreasing* relationship with proximity to EOS, in contrast to increases in actual admission, costs, and orders.

A-3.3 Required Selection on Unobservables

This appendix section details a procedure similar to that outlined in Altonji et al. (2005). The goal of this exercise is to quantify the amount of selection on unobservables necessary to explain decreases in length of stay for patients accepted at each hour near EOS. The basic intuition is that the possibility that selection on unobservables explains estimated effects can be quantified by the extents to which selection and outcomes can be explained by observables.

A-3.3.1 Conceptual Framework

Consider a condensed form of the outcomes regression Equation (2):

$$\begin{aligned} Y &= \sum_m \alpha_m A^m + \mathbf{\Omega}'\mathbf{\Gamma} \\ &= \sum_m \alpha_m A^m + \mathbf{W}'\mathbf{\Gamma}_W + \xi, \end{aligned} \tag{A-3.3}$$

where I omit subscripts for simplicity. $A^m \equiv \mathbf{1}\left(\left\lfloor t - \bar{t}\,(j,t)\right\rfloor = m\right)$ is an abbreviation for the familiar indicator for whether the time t that patient i was assigned to physician j was in the m^{th} hour from j's EOS. α_m is the causal effect of a patient being assigned in the m^{th} hour prior to EOS. $\mathbf{\Omega}$ is the full set of other variables, both observed and unobserved, that determine outcome Y, while $\mathbf{W}$ includes only observed patient, time, and provider characteristics (to be distinguished from $\mathbf{X}_{it}$ in Equation (2), which only includes patient characteristics). $\mathbf{\Gamma}$ is the causal effect of $\mathbf{\Omega}$ on Y. $\mathbf{\Gamma}_W$ is the subvector of $\mathbf{\Gamma}$ that corresponds to $\mathbf{W}$ within $\mathbf{\Omega}$, and ξ is an index of the unobserved variables.

Since variables in $\mathbf{W}$ are likely correlated with ξ, rewrite Equation (A-3.3) as

$$Y = \sum_m \alpha_m A^m + \mathbf{W}'\gamma_W + \varepsilon, \tag{A-3.4}$$

where γ_X and ε are constructed so $\text{Cov}\,(\varepsilon, \mathbf{W}) = 0$ *by definition*. Thus γ_W captures both the causal effect of $\mathbf{W}$ on Y ($\mathbf{\Gamma}_W$), as well as the portion of ξ that may be correlated with $\mathbf{W}$. Note that, for the regression estimate of α_m to be unbiased, the standard OLS assumption is that $\text{Cov}\,(\varepsilon, A^m) = 0$, or $E\,[\varepsilon\,|A^m = 1] - E\,[\varepsilon\,|A^m = 0] = 0$.

A-3.3.2 Measure of Selection on Unobservables

Altonji et al. (2005) argue for upper bound of selection on unobservables, specified by

$$\frac{E\,[\varepsilon\,|A^m = 1] - E\,[\varepsilon\,|A^m = 0]}{\text{Var}(\varepsilon)} = \frac{E\,[\mathbf{W}'\gamma_W\,|A^m = 1] - E\,[\mathbf{W}'\gamma_W\,|A^m = 0]}{\text{Var}(\mathbf{W}'\gamma_W)}, \tag{A-3.5}$$

which states that the relationship between the index of unobservables in Equation (A-3.4) and the indicator for selection A^m is equal in magnitude to the relationship between unobservable predictors of Y and A^m, respectively normalizing for variance.

They argue that this condition represents an upper bound because of observed variables are not randomly collected but rather represent characteristics that are collected precisely because they are more important for outcomes of interest. Furthermore, because many observed variables are in fact collected after the selection event, they include random shocks that cannot have influenced the selection event. This latter argument is related to the fact that I observe a rich set of patient characteristics that are either determined by the physician after accepting the

patient or are rarely observable by the physician at the time of acceptance.

A-3.3.3 Estimation of Potential Bias

In order to estimate the potential bias at the upper bound implied by Equation (A-3.5), consider the following linear selection equation:

$$A^m = \mathbf{W}'\beta_W^m + \tilde{A}^m, \tag{A-3.6}$$

where $\tilde{A}^m$ is a residual that is orthogonal to $\mathbf{W}$. Then Equation (A-3.4) can be stated as

$$Y = \sum_m \alpha_m \tilde{A}^m + \mathbf{W}'\left(\gamma_W + \sum_m \alpha_m \beta_W^m\right) + \varepsilon.$$

This leads to a statement of the potential bias due to selection on unobservables:

$$\begin{aligned} \operatorname{plim} \hat{\alpha}_m &\approx \alpha_m + \frac{\operatorname{Cov}\left(\tilde{A}^m, \varepsilon\right)}{\operatorname{Var}\left(\tilde{A}^m\right)} \\ &= \alpha_m + \frac{\operatorname{Var}\left(A^m\right)}{\operatorname{Var}\left(\tilde{A}^m\right)} \left(E\left[\varepsilon \,|A^m = 1\right] - E\left[\varepsilon \,|A^m = 0\right]\right), \end{aligned}$$

From Equation (A-3.5), the bias can be stated in terms of $E\left[\mathbf{W}'\gamma_W \,|A^m = 1\right] - E\left[\mathbf{W}'\gamma_W \,|A^m = 0\right]$:

$$\text{Bias} = \frac{\operatorname{Var}\left(A^m\right)\operatorname{Var}\left(\varepsilon\right)}{\operatorname{Var}\left(\tilde{A}^m\right)\operatorname{Var}\left(\mathbf{W}'\gamma_W\right)} \left(E\left[\mathbf{W}'\gamma_W \,|A^m = 1\right] - E\left[\mathbf{W}'\gamma_W \,|A^m = 0\right]\right) \tag{A-3.7}$$

Under the null hypothesis that $\alpha_m = 0$, γ_W can be consistently estimated by Equation (A-3.3).

I can then arrive at a consistent estimate of bias in Equation (A-3.7) with the following procedure, with results shown in Table A-3.1: For each $m \in \{-6, \ldots, -1\}$, I define $A^m \equiv \mathbf{1}\left(\left\lfloor t - \bar{t}\left(j,t\right)\right\rfloor = m\right)$ over all observations and empirically calculate $\widehat{\operatorname{Var}}\left(A^m\right)$ for each m. I also calculate $\widehat{\operatorname{Var}}\left(\tilde{A}^m\right)$ after estimating Equation (A-3.6) for each m. Similarly, I estimate $\widehat{\operatorname{Var}}\left(\varepsilon\right) = 0.160$ and $\widehat{\operatorname{Var}}\left(\mathbf{W}'\gamma_W\right) = 0.580$ from Equation (A-3.4). Equation (A-3.4) also allows me to form an estimate of selection on observables, $\hat{E}\left[\mathbf{W}'\gamma_W \,|A^m = 1\right] - \hat{E}\left[\mathbf{W}'\gamma_W \,|A^m = 0\right]$, for each m. Using the Altonji et al. (2005) condition in Equation (A-3.5) that normalized selection on unobservables is bounded by normalized selection on observables, I then calculate an upper bound of the bias due to selection on unobservables with Equation (A-3.7). As shown in Table A-3.1, the upper bound of the bias in $\hat{\alpha}_{-1}$, the effect of arriving in the last hour of shift on the length of stay, estimated by Equation (2), is -0.00124. Given that $\hat{\alpha}_{-1} = -0.5873$, this implies that normalized selection on unobservables would have to be 475 times greater than normalized selection on observables. As a comparison, in their example of the impact of Catholic school

on educational attainment, Altonji et al. (2005) argue that selection on unobservables is highly unlikely with an ratio 3.55.

A-3.4 Eliminating Selection between Physicians

This appendix section considers an additional robustness check by eliminating selection between contemporaneous physicians who could accept the same patient arriving at a given hour. Instead of using variation in the identity of the accepting physician (or more precisely, the time within that physician's shift), I only use variation in the overall composition of ED shifts at the patient's time of arrival.

The intuition behind this approach is that it, although patients may be assigned by the triage nurse or chosen by physicians as a margin of selection, it is less likely for patients to arrive at the ED at different times specifically related to the timing of shifts. Because I control for hour of the day, day of the week, and month-year interactions, correlations between patient arrival and underlying ED shift structure would have to be conditional on these time categories.

This approach is closely related to one used by Chetty et al. (2014). First, I estimate "leave-shift-out" (jackknife) EOS effects specific to shift s, using Equation (2) on all observations except those corresponding to s (Jacob et al., 2010). I denote these estimates as $\{\hat{\alpha}_{ms}\}$. This is method allows us to exclude idiosyncratic (but not systematic) shocks, including selection, on both length of stay and the right-handside of Equation (2), that would otherwise introduce bias into $\{\hat{\alpha}_{ms}\}$. Next, I construct hourly patient-weighted averages (at the level of the entire ED) that represent the overall ED shift environment. That is, for patients i arriving at time $t_i^a = t$, where t is defined at an hourly level, construct the average EOS effect

$$Q_t \equiv \frac{\sum_i \mathbf{1}\left(t_i^a = t\right) \sum_{m=-6}^{-1} \hat{\alpha}_{ms} \mathbf{1}\left(\left\lfloor t - \bar{t}\left(J\left(i,t\right),t\right)\right\rfloor = m\right) \mathbf{1}\left(S\left(J\left(i,t\right),t\right) = s\right)}{\sum_i \mathbf{1}\left(t_i^a = t\right)}, \tag{A-3.8}$$

where $J\left(i,t\right)$ is a physician assignment function for patient i at time t, and $S\left(j,t\right)$ is a shift assignment function for physician j at time t.

A-3.4.1 Unobservable Selection between Physicians within Hour

I evaluate systematic bias of $\{\hat{\alpha}_{ms}\}$ due to selection (on unobservables) between physician within hour. To do this, I average out within-hour selection by constructing patient-weighted averages Y_t of (residualized) length of stay $\tilde{Y}_{ijkpt}$:

$$Y_t \equiv \frac{\sum_i \mathbf{1}\left(t_i^a = t\right) \tilde{Y}_{ijkpt}}{\sum_i \mathbf{1}\left(t_i^a = t\right)} \tag{A-3.9}$$

where

$$\tilde{Y}_{ijkpt} \equiv Y_{ijkpt} - \left[\sum_m \hat{\gamma}_m \mathbf{1} \left(\lfloor t - \underline{t}\left(j,t\right) \rfloor = m \right) + \mathbf{X}'_{it}\hat{\beta} + \mathbf{T}'_t\hat{\eta} + \hat{\zeta}_p + \hat{\nu}_k \right]. \tag{A-3.10}$$

Coefficients $\hat{\gamma}_m$, $\hat{\beta}$, $\hat{\eta}$, $\hat{\zeta}_p$, and $\hat{\nu}_k$ are estimated using within-EOS variation from an equation very similar to Equation (2):

$$\begin{aligned} Y_{ijkpt} \quad = \quad & \sum_{m=-6}^{-1} \alpha_m \mathbf{1} \left(\left\lfloor t - \bar{t}\left(j,t\right) \right\rfloor = m \right) + \sum_m \gamma_m \mathbf{1} \left(\lfloor t - \underline{t}\left(j,t\right) \rfloor = m \right) + \\ & \mathbf{X}'_{it}\beta + \mathbf{T}'_t\eta + \zeta_p + \nu_k + \varepsilon_{ijkpt}, \end{aligned}$$

where I use physician fixed effect ν_k instead of physician-team fixed effects ν_{jk} to broaden the number of observations for which I observe an identified residual. This approach, which includes effects for time to EOS, only uses within-EOS-time variation to estimate coefficients and therefore provides consistent estimates even if the covariates are correlated with time relative to EOS.

The regression

$$Y_t = a + bQ_t + \chi_t, \tag{A-3.11}$$

quantifies the degree of "forecast bias" due to systematic selection of patients arriving within t across physicians,

$$B\left(\hat{\alpha}_{ms}\right) = \mathrm{Cov}\left(\varepsilon_{ijkpt}, \hat{\alpha}_{ms}\right) / \mathrm{Var}\left(\hat{\alpha}_{ms}\right), \tag{A-3.12}$$

for $\lfloor t - \underline{t}\left(j,t\right) \rfloor = m$ and $S\left(j,t\right) = s$. $B\left(\hat{\alpha}_{ms}\right) = 1 - b$, under the assumption that

$$\mathrm{Cov}\left(A_t, \chi_t\right) = 0. \tag{A-3.13}$$

This assumption states that there is no selection of unobservable patient types across ED times conditional on time categories (i.e., hour of the day, day of the week, and month-year interactions). This assumption is much more plausible than the baseline assumption that there is no selection of unobservable patient types arriving at the same time across physicians.

Column 1 of Table A-3.2 reports estimates of b from Equation (A-3.11). The point estimate of b is 1.029 with a robust standard error of 0.060 (clustered at each hour of t), which reflects tight estimation indistinguishable from 1 (i.e., I cannot reject the hypothesis of $B\left(\hat{\alpha}_{ms}\right) = 0$). That is, under the assumption of no selection of unobservable patient types across ED times, I cannot show that there is bias caused by selection of unobservable patient types arriving at the same time across physicians. Panel A of Figure (6) plots the relationship between Y_t and Q_t nonparametrically, dividing the data into 20 equal-sized groups ("vigintiles") according to Q_t. This plot nonparametrically represents the conditional expectation function of Y_t conditional on Q_t. The relationship is highly linear, with slope close to 1.

A-3.4.2 Observable Selection between Hours

I use a similar exercise to consider the amount of selection on observables across hours, conditional on time categories. Similar to the analysis in Appendix A-3.2, I consider two sets of *ex post* observable characteristics, $\mathbf{X}_{it}^{prior}$ and $\mathbf{X}_{it}^{full}$, to form predictions about length of stay. The former set includes characteristics that are observable to the physician prior to acceptance, while the latter set is a superset that also includes characteristics that generally are not observable to the physician until after acceptance. However, as with Y_t, I average predictions for all patients within a given hour, again eliminating selection across physicians within hour. This exercise therefore evaluates the degree of selection remaining across hours (on characteristics nevertheless controlled for in estimating EOS effects).

For each variable in $\mathbf{X}_{it}^{full}$, I form residualized variables obtained after subtracting predictions of each variable based on time categories, $\mathbf{T}_t$, and indicators for hours relative to shift beginning, $\lfloor t - \underline{t}(j,t) \rfloor$. Using sets of residualized characteristics, $\tilde{\mathbf{X}}_{it}^{prior}$ and $\tilde{\mathbf{X}}_{it}^{full}$, I construct predictions $\hat{Y}_{ijkpt}^{prior}$ and $\hat{Y}_{ijkpt}^{full}$. Similar to Equation (A-3.9), I average these predictions over all patients arriving at a given hour:

$$\hat{Y}_t^{set} \equiv \frac{\sum_i \mathbf{1}\left(t_i^a = t\right) \hat{Y}_{ijkpt}^{set}}{\sum_i \mathbf{1}\left(t_i^a = t\right)}, \tag{A-3.14}$$

where t denotes an hour, and $set \in \{prior, full\}$.

The regression

$$\hat{Y}_t^{set} = a + b^{set} Q_t + \chi_t \tag{A-3.15}$$

quantifies the degree of selection across hours, as predicted by characteristics $\mathbf{X}_{it}^{set}$: Under (A-3.13) , $b^{set} = \text{Cov}\left(\hat{\alpha}_{ms}, \hat{Y}_t^{set}\right) / \text{Var}\left(\hat{\alpha}_{ms}\right)$ for $\lfloor t - \underline{t}(j,t) \rfloor = m$ and $S(j,t) = s$. Although the assumption in Equation (A-3.13) is not directly testable, a lack of observable selection (b^{set} is indistinguishable from 0) supports this assumption.

Columns 2 and 3 of Table A-3.2 report of estimates b^{prior} and b^{full}, respectively, from Equation A-3.15. Both estimates are small and indistinguishable from 0: The point estimate of $b^{prior}$is 0.029 (robust standard error 0.025), and the point estimate of b^{full} is 0.024 (robust standard error 0.026). Panels B and C of Figure (6) show corresponding nonparametric expectations of $\hat{Y}_t^{prior}$ and $\hat{Y}_t^{full}$, respectively conditional on Q_t, where the data is again divided into vigintiles of Q_t. The relationship is again linear, but consistent with the regression results, there is no relationship between length of stay predicted by time relative to EOS (Q_t) and that predicted by patient characteristics. I consider this as strong evidence supporting (A-3.13).

A-4 Structural Model Implementation

This appendix details the procedure to simulate observations under counterfactual assignment policies and impute overall costs under these policies, as discussed more briefly in Section 7.

To summarize, I first estimate arrival and discharge functions in discrete time. Actual arrivals and discharges imply patient censuses, or the number of patients under the care of a physician at each point in discrete time, i.e., those who have arrived and have not yet been discharged. I create counterfactual assignment policies by modifying time in the arrival function, and I simulate of patient observations (including arrival times, discharge times, and patient censuses) using this modified arrival function and the (unmodified) discharge function. Finally, I impute of overall costs by regressing simulated workload-adjusted length of stay on time relative to EOS and translating decreases in workload-adjusted length of stay into increases in resource-utilization costs. Overall costs are the sum of costs due to patient care and physician time.

A-4.1 Estimating Arrival and Discharge Functions

In each fifteen-minute time interval of each of the 23,990 shifts during the study period ranging from June 2005 to December 2012, I calculate the number of patients assigned to the physician on shift s during time interval t. Restricting to $t \in \left[\underline{t}(s) - 3 \text{ hours}, \overline{t}(s)\right]$ yields 1,151,888 observations over t and s with patient arrival (or "assignment") numbers

$$N_{st} \equiv \sum_i \mathbf{1}\left(t_i^a = t\right) \mathbf{1}\left(S\left(J\left(i\right), t_i^a\right) = s\right),$$

where t_i^a is the arrival time of patient i, $S\left(j,t\right)$ is a shift assignment function for physician j and time t, and $J\left(i\right)$ is a physician assignment function, assuming for notational convenience that each patient has only one visit.

Of $\sum_{s,t} N_{st} = 370,843$ patients arriving during valid times, I further restrict the estimation sample to arrivals and discharges of 350,053 patients whose length of stay is at most twelve hours and who arrived at most twelve hours prior to EOS. The remaining 20,790 patients, whom I denote as $i \in I^{\text{outside}}$, are therefore not modeled in either arrivals or discharges, but I count them toward workload defined below. As I describe in Section A-4.2, I take arrivals and discharges of patients $i \in I^{\text{outside}}$ as fixed in every simulation.

For N_{st}, I estimate a zero-inflated Poisson model. I call this an arrival or assignment function $\mathcal{A}\left(t, \sigma_s, w_{j,t-1}\right)$, which depends on t, shift characteristics σ_s of s (i.e., shift type $\langle \ell, \underline{o}, \overline{o} \rangle_s$ and time of EOS $\overline{t}\left(s\right)$), and physician j's census (or workload) $w_{j,t-1}$ in the previous period (for j satisfying $s\left(j,t\right) = s$). w_{jt} is defined in Equation (3), which I slightly rephrase here as

$$w_{jt} \equiv \sum_i \mathbf{1}\left(t_i^a \geq t\right) \mathbf{1}\left(t \leq t_i^d\right) \mathbf{1}\left(j = J\left(i\right)\right), \tag{A-4.16}$$

where t_i^d is the corresponding discharge order time for patient i, again taking advantage of the notational assumption that i refers to a unique patient visit.

As introduced by Lambert (1992), the zero-inflated Poisson model allows for two "regimes," one which always yields $N_{st} = 0$, and the second which follows a Poisson process leading to

$N_{st} \geq 0$. In particular,

$$\begin{aligned}
\Pr\left(N_{st}=0\right) &= \Pr\left(\text{Regime 1}\right)+\Pr\left(N_{st}=0|\,\text{Regime 2}\right)\Pr\left(\text{Regime 2}\right); \\
\Pr\left(N_{st}=n\right) &= \Pr\left(N_{st}=n|\,\text{Regime 2}\right)\Pr\left(\text{Regime 2}\right),\ n=1,2,\ldots,
\end{aligned}$$

where $\Pr\left(\text{Regime 1}\right)+\Pr\left(\text{Regime 2}\right)=1$. In standard fashion, $\Pr\left(\text{Regime 1}\,|t,\sigma_s\right)$ is specified as the logistic function of a linear combination of shift-type indicators interacted with splines of $\bar{t}_s-t$. $\Pr\left(N_{st}=n|\,\text{Regime 2}\right)$ is specified as a Poisson model:

$$\Pr\left(N_{st}=n|\,\text{Regime 2}\right)=\frac{\exp\left(-\lambda_{st}\right)\lambda_{st}^{n}}{n!},\ n=0,1,2,\ldots,$$

where $\log\lambda_{st}=\log E\left[N_{st}|\,t,\sigma_s,w_{j,t-1}\right]$ is a linear combination of splines of time of the day, splines of month-year interactions, indicators for day of the week, splines of $\bar{t}_s-t$, splines of $w_{j,t-1}$, interactions between splines of $\bar{t}_s-t$ and splines of $w_{j,t-1}$, and interactions between shift-type indicators interacted with splines of $\bar{t}_s-t$.

For discharge times t_i^d, I estimate a logit hazard model as

$$\Pr\left(t_i^d=t\,\middle|\,t_i^d\geq t-1\right)\equiv\mathcal{D}\left(t,\sigma_s,w_{jt},\hat{\tau}_{ist}\right)=\frac{1}{1+\exp\left(-h\left(t,\sigma_s,w_{jt},\hat{\tau}_{ist}\right)\right)},$$

where $h\left(t,\sigma_s,w_{jt},\hat{\tau}_{ist}\right)$ is estimated, separately for shifts of different length ℓ, as a linear combination of indicators for hour of the day of t_i^a, splines of $t-t_i^a$, splines of $\bar{t}\left(s\right)-t$, splines of $\bar{t}\left(s\right)-t_i^a$, splines of w_{jt}, predicted log length of stay ($\hat{\tau}_{ist}$), interactions between splines of $t-t_i^a$ and splines of $\bar{t}\left(s\right)-t_i^a$, interactions between splines of $\bar{t}\left(s\right)-t$ and splines of w_{jt}, and interactions between $\hat{\tau}_{ist}$ and splines of $\bar{t}\left(s\right)-t_i^a$.

Predicted log length of stay ($\hat{\tau}_{ist}$) is a linear combination of indicators for day of the week, month-year interactions, patient age and squared age, sex, ESI indicators, Elixhauser indicators, race indicators, language indicators, pod indicators, and physician-nurse-resident joint indicators. As a separate model, I estimate a model $\tilde{\tau}_{ist}$ of $\hat{\tau}_{ist}$ based on day of the week, month-year interactions, pod indicators, shift-type indicators interacted with splines of $\bar{t}\left(s\right)-t$, and physician indicators. The reason for the first prediction ($\hat{\tau}_{ist}$ of τ_{ijkpt}) is to condense a large number of characteristics about the patient visit into a single linear score for ease of estimation. The reason for the second prediction ($\tilde{\tau}_{ist}$ of $\hat{\tau}_{ist}$) is to allow simulation of predicted log length of stay without simulating many of the characteristics that enter into $\hat{\tau}_{ist}$, such as nurse and resident identities, which I discuss in the next section.

A-4.2 Simulating Patient Observations

After estimating arrival and discharge functions from actual data, I simulate patient arrivals and discharges under counterfactual assignment policies. I create the counterfactual policies, $\mathcal{A}_{\Delta}\left(t,\sigma_s,w_{j,t-1}\right)\equiv\mathcal{A}\left(\breve{t}\left(t,s,\Delta\right),\sigma_s,w_{j,t-1}\right)$, by modifying patient assignment indexed by a time

shift Δ, so that patients are assigned as if time were $\breve{t}(t, s, \Delta)$ rather than t. Starting at $|\Delta|$ hours before EOS, assignments are "curtailed" assignments with $\breve{t} > t$ if $\Delta < 0$ or "extended" with $\breve{t} < t$ if $\Delta > 0$. Formally,

$$\breve{t}(t, s, \Delta) = \begin{cases} t + \kappa \max\left(1 + t - \bar{t}(s), \Delta\right), & \Delta < 0 \\ t, & \Delta = 0 \\ t + \max\left(0, \min\left(\Delta + t - \bar{t}(s), \Delta\right)\right), & \Delta > 0 \end{cases}, \tag{A-4.17}$$

where $\kappa = \max\left(0, \min\left(1, t - \bar{t}(s) - \Delta\right)\right) \in [0, 1]$ is a scale to ensure that $\breve{t}$ is continuous in t when $\Delta < 0$. Figure 8 shows example counterfactual policies. While assignments are explicitly modified in these policies, parameters of the underlying discharge function is unchanged and remains under the control of the physicians. Discharge behavior of course responds to w_{jt}, as it increases or decreases near EOS, with $\Delta > 0$ or $\Delta < 0$, respectively.

Specifically, I follow this procedure for each simulation r of counterfactual policy Δ:

1. Start t at three hours before the beginning of each shift s. Set $w_{j,t-1}^{\Delta,r} = 0$.
2. Determine new assignments at t for each s.
   (a) Simulate $N_{st}^{\Delta,r}$new assignments for s at t, using $\mathcal{A}_\Delta$. Denote each of these new assignments with an unused $i \notin S^{\text{outside}}$, note that $t_{\Delta,r,i}^{a} = t$, and simulate predicted log length of stay $\tilde{\tau}_{ist}^{\Delta,r}$.
   (b) Assign patients $i \in I^{\text{outside}}$ where $t_i^{\text{outside},a} = t$ to the relevant shifts s.
3. Calculate workload $w_{jt}^{\Delta,r}$ by Equation (A-4.16).
4. If $t \geq \underline{t}(s)$ and $w_{jt}^{\Delta,r} > 0$, determine discharges at t for each s.
   (a) Simulate $d_{it}^{\Delta,r} \equiv \mathbf{1}\left(t_{\Delta,r,i}^{d} = t\right)$ for each $i \notin S^{\text{outside}}$ where $d_{i,t-1}^{\Delta,r} = 0$, using $\mathcal{D}$.
   (b) Discharge patients $i \in S^{\text{outside}}$ where $t_i^{\text{outside},d} = t$ from the relevant shifts s.
5. The procedure is complete for s such that $t \geq \bar{t}(s)$ and $w_{jt}^{\Delta,r} = 0$. For the remaining s, revise $t = t + 1$ and return to Step #2.

The resulting collection of $t_{\Delta,r,i}^{a}$, $t_{\Delta,r,i}^{d}$, and $w_{jt}^{\Delta,r}$ form the underlying simulated data. Simulated workload-adjusted length of stay for patient i under physician j can be calculated by dividing i's simulated length of stay by simulated average censuses under j during i's length of stay. Slightly

adapting Equation (5) to discrete time,

$$\begin{aligned}\tilde{Y}_{ij}^{\Delta,r} &\equiv \tau_i^{\Delta,r}/\overline{w}_{ij}^{\Delta,r}, \\ &= 0.25 \cdot \max\left(t^d_{\Delta,r,i} - t^a_{\Delta,r,i}, 0.3\right) \left[\frac{1}{t^d_{\Delta,r,i} - t^a_{\Delta,r,i} + 1} \sum_{\tilde{t}=t^a_{\Delta,r,i}}^{t^d_{\Delta,r,i}} w_{j\tilde{t}}^{\Delta,r}\right]^{-1}. \quad \text{(A-4.18)}\end{aligned}$$

The term $0.25 \cdot \max\left(t^d_{\Delta,r,i} - t^a_{\Delta,r,i}, 0.3\right)$ reflects that, in actual data, $\hat{E}[\tau_i] \approx 0.075$ hours if $t_i^d = t_i^a$, but otherwise $\hat{E}[\tau_i] \approx 0.25\left(t_i^d - t_i^a\right)$ hours if $t_i^d > t_i^a$ (recall that t is in fifteen-minute intervals).

A-4.3 Imputing Costs

Having simulated arrival and discharge data, I am now in the position to impute overall costs for each counterfactual simulation r of Δ. Overall costs include physician-time, patient-time, and hospital-resource costs. Repeating Equation (6):

$$\text{Costs}_\Delta^r = \text{PhysicianTime}_\Delta^r + \text{PatientTime}_\Delta^r + \text{HospitalResources}_\Delta^r. \quad \text{(A-4.19)}$$

The first cost, physician-time costs, represents the value of leisure foregone. Physician hours can increase either if a peer must arrive earlier before the index physicians EOS, or if the index physician must stay longer past EOS:

$$\text{PhysicianTime}_\Delta^r = \text{Wage} \times \sum_s \left(\text{WorkCompletionTime}_\Delta^{s,r} - \text{PeerArrivalTime}_\Delta^{s,r}\right).$$

"Slacking off" in the assignment policy occurs earlier relative to EOS mechanically requires peers to arrive earlier. In the actual data, there are generally two unseen patients at the time of peer arrival (see Figure A-5.3). I therefore model $\text{PeerArrivalTime}_\Delta^{s,r}$ as when there are two unseen patients near $\bar{t}(s)$, based on the assignment policy and an exogenous pod flow rate of 2.22 patients per hour (see Figure 3). I model $\text{WorkCompletionTime}_\Delta^{s,r}$ (when the physician on shift s leaves the ED) as the time between when all but one or two patients have been discharged. This empirically matches the stated work completion time of generally two to three hours past EOS, although results are insensitive to the precise definition of work completion. Implicit in this rule is that physicians are not more likely to pass off patients with more work at EOS; given that work completion time is really insensitive to being assigned more work at EOS (due to quicker discharges), this is unlikely to be quantitatively important. I multiply physician-hours by a base-case wage of \$120 per hour but also consider the extreme case of \$600 per hour.

The second cost, patient-time costs, reflects the value of patient time:

$$\text{PatientTime}_{\Delta}^{r} = \text{TimeValue} \times \sum_{i} \left(\tau_i^{\Delta,r} - E\left[\tau_i^{0,r} \right] \right),$$

where TimeValue = \$20/hour, or roughly the average hourly wage in the US.

The third cost in Equation (A-4.19), hospital-resource costs, represents resource costs, via formal utilization and admissions, incurred by the physician. As shown in Section A-2 and Table 2, workload-adjusted length of stay, formal orders, admissions, and total costs all increase only in the last hour of shift, suggesting that workload-adjusted length of stay is a good measure of time that increases patient-care costs as it is decreased. In each simulation r of each policy Δ, I estimate the EOS effect on workload-adjusted length of stay by coefficients $\hat{\alpha}_m^{\Delta,r}$ in this regression:

$$\begin{aligned} \log \tilde{Y}_{ij}^{\Delta,r} \quad = \quad & \alpha^{\Delta,r} + \sum_{m=-6}^{-1} \alpha_m^{\Delta,r} \mathbf{1}\left(\left\lfloor t_{\Delta,r,i}^{a} - \bar{t}\left(S\left(j, t_{\Delta,r,i}^{a} \right) \right) \right\rfloor = m \right) + \\ & \mathbf{g}\left(t_{\Delta,r,i}^{a} - \underline{t}\left(S\left(j, t_{\Delta,r,i}^{a} \right) \right) \right)' \gamma_g^{\Delta,r} + \varepsilon_{ij}^{\Delta,r}, \end{aligned} \tag{A-4.20}$$

where $\tilde{Y}_{ij}^{\Delta,r}$ is simulated workload-adjusted length of stay from Equation (A-4.18), and $\mathbf{g}(\cdot)$ creates a vector of cubic splines of assignment time relative to shift beginning.

In simulated data with $\Delta = 0$, I estimate $\hat{\alpha}_{-1}^{0} \equiv \frac{1}{100}\sum_{r=1}^{100} \hat{\alpha}_{-1}^{0,r} = -0.240$ and $\hat{\alpha}_{-2}^{0} \equiv \frac{1}{100}\sum_{r=1}^{100} \hat{\alpha}_{-2}^{0,r} = -0.059$, which implies that workload-adjusted length of stay decreases by 18.1% in the last hour of shift under the observed assignment policy. Note that this difference is slightly smaller (more conservative) than that implied by coefficients $\hat{\alpha}_{-1} = -0.232$ and $\hat{\alpha}_{-2} = -0.069$ estimated without simulation using actual data (Table A-5.3). Given that total costs increase by 20.8% in the last hour prior to EOS, I estimate the elasticity of hospital-resource costs to workload-adjusted length of stay, for decreases in workload-adjusted length of stay that are 5.9% below baseline, as $20.8\%/-18.1\% = -1.15$. I thus calculate hospital-resource costs as

$$\begin{aligned} \text{HospitalResources}_{\Delta}^{r} \quad = \quad & \sum_{s} \sum_{t=\underline{t}(s)}^{\bar{t}(s)} \mathbf{1}\left(\left\lfloor t - \bar{t}\left(s\right) \right\rfloor = m \right) N_{st}^{\Delta,r} \times \\ & \sum_{m} \exp\left(\text{BaseLogCosts} - 1.15 \cdot \min\left(0, \hat{\alpha}_m^{\Delta,r} - \hat{\alpha}_{-2}^{0} \right) \right), \end{aligned} \tag{A-4.21}$$

where BaseLogCosts = $[\log \$ +]\, 6.750$. Note hospital-resource costs increase with greater assignments (higher Δ) both because per–patient costs increase, and the number of patients that this applies to also increases. As discussed in the main paper, I conservatively assume no negative effects on patient health, even as physicians produce less information for the discharge decision, since I observe none in sample (Table 2).

A-4.4 Imputing the Value of Leisure

I can also impute the revealed value of leisure in terms of hospital-resource costs by calculating the ratio between extra hospital-resource costs incurred and leisure time gained as a result of the physician discharge behavior near EOS. The revealed discharge function $\mathcal{D}\left(t, \sigma_s, w_{jt}, \hat{\tau}_{ist}\right)$ increases the discharge hazard as t approaches $\bar{t}\left(s\right)$, shortening patient-adjusted length of stay and increasing hospital-resource costs. Similar to modifying t in the assignment policy $\breve{t}\left(t, s, \Delta\right)$, Equation (A-4.17), I examine what discharges would look like if not influenced by EOS behavior by modifying t in the discharge function. That is, I consider $\mathcal{D}_0\left(t, \sigma_s, w_{jt}, \hat{\tau}_{ist}\right) \equiv \mathcal{D}\left(\tilde{t}\left(t, s\right), \sigma_s, w_{jt}, \hat{\tau}_{ist}\right)$ as a modified discharge function, where

$$\tilde{t}\left(t, s\right) = \min\left(t, \bar{t}\left(s\right) - 4\right),$$

ensuring that discharge behavior does not reflect EOS behavior, even if t approaches $\bar{t}\left(s\right)$, since $\tilde{t}\left(t, s\right)$ is at least four hours before EOS.

I then evaluate differences in hospital-resource costs and work-completion time under both of these discharge functions. The ratio between these two differences reveals physicians' implicit valuation of an hour of leisure in terms of hospital-resource costs:

$$\text{LeisureValue}^r_\Delta = -\frac{\text{HospitalResources}^r_\Delta|\mathcal{D} - \text{HospitalResources}^r_\Delta|\mathcal{D}_0}{\text{WorkCompletionTime}^r_\Delta|\mathcal{D} - \text{WorkCompletionTime}^r_\Delta|\mathcal{D}_0}.$$

A-5 Additional Results

In this appendix, I present the following additional empirical results, as well as a brief discussion of some of these results:

- Table A-5.1 describes the process of constructing the sample, including the number of observations in each step.
- Table A-5.2 lists the number of observations for each shift type. Observations are counted in terms of unique shifts, hours, potential patients (who could be assigned to a shift of that shift type at time of arrival), and actual patients (who are assigned to a shift of that shift type).
- Table A-5.3 reports coefficients for EOS effects on workload-adjusted length of stay, as a continuation of Table 3. Results in this table are estimated with the full set of controls but only control for time relative to shift beginning. Results are estimated on both actual and simulated data.
- Figure A-5.1 shows example weekly pod schedules.

- Figure A-5.2 shows evidence on how long physicians stay past EOS in terms of when three types of orders are written: the last order by the attending physician of record (AOR), the first (non-resident) physician order after the AOR's orders, and the last discharge order.
- Figure A-5.3 shows average patient counts ("censuses") for physicians in shifts with different overlap $\overline{o}$.
- Figure A-5.4 shows the fit between actual and simulated data, where the simulated assignments, length of stay, and censuses are calculated by discrete-time functions of patient assignment and discharge.
- Figure A-5.5 shows coefficients for EOS effects on workload-adjusted length of stay, reported in Table A-5.3, estimated on both actual and simulated data.

In Table A-5.3, I present evidence qualitatively consistent with results in Table 3, except that I do not control for patient characteristics, time indicators other than time relative to shift beginning, and provider identities. I consider these more parsimonious regressions in Table A-5.3 to operationalize workload-adjusted length of stay as the key substitute for hospital-resource costs in the structural model in Section 7, in which simulating the rich set of covariates would either be impractical. Specifically, as workload-adjusted length of stay decreases, hospital-resource costs increase via increased formal utilization and admission likelihood. Tables 3 and A-5.3 both show that workload-adjusted length of stay *increases* above baseline in the hours before the last of shift (while utilization is unchanged). This could be consistent with "foot-dragging" in which physicians delay discharge but do not otherwise change patient care Chan (2015); in the structural model, I therefore assume that increases in workload-adjusted length of stay do not change hospital-resource costs.

Figure A-5.2 shows the cumulative densities of orders relevant to physicians staying past EOS. I do not observe when physicians actually go home, but I can create some relevant bounds based on three types of orders: the last order by the attending physician of record (AOR), the first (non-resident) physician order after the AOR's orders, and the last discharge order. As described in the figure notes, most visits do not have orders written by physicians (i.e., most visits have orders solely written by resident physicians). Cumulative densities for physician orders are calculated using visits in which the relevant physician order is observed at least once. Physicians obviously must be present to write their own orders, but the last AOR order is a definite lower bound, since most visits do not have AOR orders, and work such as dictating patient notes does not require orders. Orders by the next non-resident physician are definitive proof that another physician has started to assume the care of some of the index physicians' original patients, although it is still possible that the physician may be staying to finish work on other patients, including work that does not require orders. The last discharge order is likely an overestimate of the time spent past EOS, since some patients may be passed off or have well-defined discharge plans before the discharge order.

Figure A-5.4 shows the fit between actual data and simulated data from the structural model described in Sections 7 and A-4. Assignments (Panel A) are simulated according to a zero-inflated Poisson model, and length of stay (Panel B) results from discharges simulated according to a logit hazard model, both of which consider time in fifteen-minute discrete intervals. Censuses (Panel C) – or workload as measured by numbers of patients under a physician's care (between assignment and discharge) – are simulated by both the assignment and discharge models. Workload-adjusted length of stay, calculated by Equation (5), is derived from discharges and censuses. Figure A-5.5 shows the fit in regression coefficients (also reported in Table A-5.3) of workload-adjusted length of stay between actual and simulated data.

Table A-3.1: Potential Bias from Selection on Unobservables

	$\widehat{\text{Var}}(A^m)$	$\widehat{\text{Var}}\left(\tilde{A}^m\right)$	Selection on observables	Bias upper bound	$\hat{\alpha}_m$	$\hat{\alpha}_m$ as bias multiple
Patient selection into hour prior to EOS (A^m)						
Last hour (A^{-1})	0.00249	0.00062	-0.00111	-0.00124	-0.5873	474.93
Second hour (A^{-2})	0.02658	0.00387	-0.01103	-0.02086	-0.2869	13.75
Third hour (A^{-3})	0.07442	0.00784	0.00223	0.00584	-0.1230	-21.05
Fourth hour (A^{-4})	0.08956	0.01053	-0.00381	-0.00893	-0.0907	10.16
Fifth hour (A^{-5})	0.10191	0.01287	-0.03295	-0.07192	-0.0232	0.32
Sixth hour (A^{-6})	0.10851	0.01391	-0.04192	-0.09014	-0.0103	0.11

Note: This table reports estimates in a procedure based on Altonji et al. (2005) to calculate potential bias from selection on unobservables, as described in Appendix A-3.3. Selection is modeled for whether a patient is assigned in the m^{th} hour prior to EOS (A^m) by Equation (A-3.6), the residual of which is $\tilde{A}^m$. Selection on observables is defined as $\hat{E}\left[\mathbf{W}'\gamma_W \,|A^m = 1\right] - \hat{E}\left[\mathbf{W}'\gamma_W \,|A^m = 0\right]$, where γ_W is estimated from Equation (A-3.4). Using the condition from Altonji et al. (2005), in Equation (A-3.5), which states that normalized selection on unobservables is at most equal in magnitude to normalized selection on observables, an upper bound of bias from selection on unobservables is calculated from Equation (A-3.7). I use $\widehat{\text{Var}}(\varepsilon) = 0.160$ and $\widehat{\text{Var}}(\mathbf{W}'\gamma_W) = 0.580$ in this calculation. $\hat{\alpha}_m$ is estimated by Equation (2); for convenience, results are repeated from the last column of Table 1. Finally $\hat{\alpha}_m$ is stated as a multiple of the bias upper bound in the last column of this table.

Table A-3.2: Mean EOS Effect on Mean Actual and Predicted Log Length of Stay

	(1)	(2)	(3)
	Mean actual, Y_t	Mean predicted, $\hat{Y}_t^{prior}$	Mean predicted, $\hat{Y}_t^{full}$
Mean EOS effect, Q_t	1.029***	0.029	0.024
	(0.060)	(0.025)	(0.025)
Number of visits	409,352	409,352	409,352
Number of shifts	22,501	22,501	22,501
Number of hour cells	63,345	63,355	63,355

Note: This table reports coefficient estimates and standard errors in parentheses for Equation (A-3.11) in Column 1 and for Equation (A-3.15) in Columns 2 and 3. Predicted and actual log lengths of stay are all averaged within hour cell and weighted by visit. The key independent variable is the log length of stay predicted by the times to EOS, defined by Equation (A-3.8) as Q_t. $Q_t$is calculated as follows: First, coefficients on time relative to EOS are calculated from (2) using a leave-shift-out sampling. Next, these coefficients are averaged across shifts in process at hour t, weighted by visits. I calculate the dependent variable for Column 1 as follows: I calculate residualized actual log length of stay, by subtracting expected log length of stay based on all covariates listed in the note for Table 1, using only variation within time to EOS. To calculate predicted log length of stay by patient characteristics (Columns 2 and 3), I residualize the characteristics by time categories and use within-EOS-time variation to predict log length of stay. Patient characteristics and time categories are described in the notes for Figure 5 and Table 1, respectively. IOLS is performed keeping visits as observations, but standard errors are clustered by the hour of patient arrival. * denotes significance at 10% level, ** denotes significance at 5% level, and *** denotes significance at 1% level.

Table A-5.1: Sample Definition

Sample description or step	Variables added	Observations
1. Raw visit data	Patient demographics, clinical diagnoses, process times (arrival at ED, arrival at bed, discharge order, discharge with destination), treatment pod, 30-day mortality, providers of record (physician, resident or physician assistant, nurse)	442,244
2. Drop visits with patients leaving before being assigned by physician or discharged		426,899
3. Merge with physician order data and bed audit data	Detailed physician orders with timestamps for medication, intravenous fluids, laboratory tests, radiology tests, and nursing orders; timestamps for bed movements	411,198
4. Merge with pod schedules	Shift types, start times, end times, and managerial locations	398,563
5. Identify visits with physician of record in visit data matching with schedules		372,224

Note: This table describes each step in sample construction. Variables included in each step are listed in the second column, and the number of observations resulting from each step are in the third column.

Table A-5.2: Shift Type Observation Numbers

Shift type	Shifts	Hours	Potential patients	Actual patients
$\langle 7,0,1\rangle$	95	665	1,645	1,160
$\langle 7,1,0\rangle$	237	1,659	6,674	2,597
$\langle 7,1,1\rangle$	101	707	4,281	1,783
$\langle 8,0,1\rangle$	319	2,552	8,453	4,952
$\langle 8,1,0\rangle$	174	1,392	7,440	1,981
$\langle 9,0,1\rangle$	3,453	30,879	84,292	58,589
$\langle 9,0,2\rangle$	325	2,349	6,411	4,541
$\langle 9,0,4\rangle$	408	2,898	9,326	4,839
$\langle 9,0,6\rangle$	364	3,276	16,186	5,899
$\langle 9,1,0\rangle$	3,414	30,528	118,030	59,897
$\langle 9,1,1\rangle$	2,909	26,181	116,108	54,221
$\langle 9,1,4\rangle$	2,249	19,170	80,279	28,694
$\langle 9,1,5\rangle$	60	540	2,554	892
$\langle 9,1,6\rangle$	211	1,899	8,157	2,524
$\langle 9,2,0\rangle$	464	3,294	12,027	6,317
$\langle 9,3,1\rangle$	485	3,277	17,013	6,699
$\langle 9,3,3\rangle$	60	540	3,226	1,089
$\langle 9,4,0\rangle$	347	2,347	9,996	3,994
$\langle 9,4,1\rangle$	212	1,908	8,974	3,370
$\langle 9,4,3\rangle$	426	2,752	16,730	5,344
$\langle 9,4,4\rangle$	772	5,094	26,094	9,413
$\langle 9,4,6\rangle$	2,141	19,269	99,726	29,007
$\langle 9,5,3\rangle$	60	540	2,851	1,043
$\langle 9,6,0\rangle$	634	5,706	34,943	9,244
$\langle 9,6,1\rangle$	1,504	13,536	61,197	21,861
$\langle 9,6,4\rangle$	575	5,175	31,088	9,597
$\langle 9,9,1\rangle$	353	3,177	15,965	4,598
$\langle 10,0,0\rangle$	176	1,760	4,812	2,578
$\langle 10,0,1\rangle$	243	2,430	5,783	4,615
$\langle 10,0,2\rangle$	137	1,040	2,631	1,901
$\langle 10,0,4\rangle$	139	1,050	3,616	2,378
$\langle 10,1,0\rangle$	277	2,770	9,092	4,401
$\langle 10,4,0\rangle$	139	1,050	4,335	1,834
$\langle 12,0,0\rangle$	142	1,704	4,119	2,423
$\langle 12,4,9\rangle$	319	3,828	16,490	5,566
Total	23,924	206,942	860,544	369,841

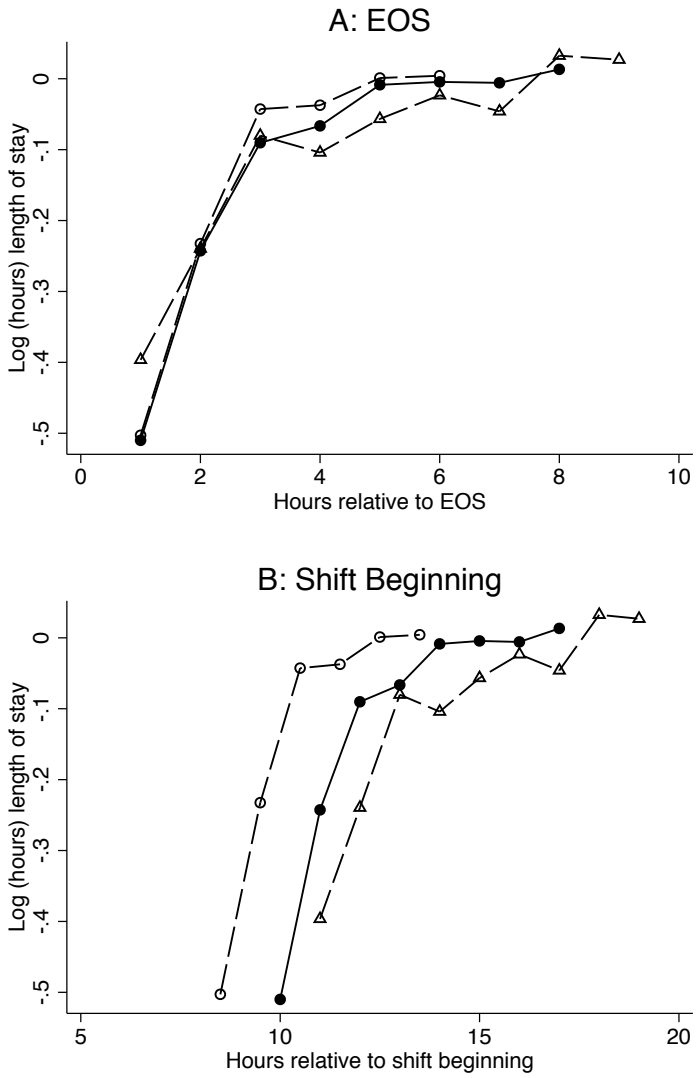
Note: This table lists the number of observations for each shift type, each defined as $\langle \ell, \underline{o}, \overline{o}\rangle$, where ℓ is the shift length in hours, $\underline{o}$ is the overlap in hours with a previous shift, and $\overline{o}$ is the overlap in hours with a subsequent shift in the same location. Observations are counted in terms of unique shifts, hours, potential patients (patients who arrive at the ED during a time when there is a shift of type $\langle \ell, \underline{o}, \overline{o}\rangle$ in progress), and actual patients (patients who are treated by a physician on a shift of type $\langle \ell, \underline{o}, \overline{o}\rangle$).

Table A-5.3: Effect on Workload-adjusted Length of Stay by Shift Overlap

	(1)	(2)	(3)	(4)	(5)	(6)
	All $\overline{o}$	$\overline{o} \leq 1$	$\overline{o} \geq 2$	All $\overline{o}$	$\overline{o} \leq 1$	$\overline{o} \geq 2$
Hour prior to EOS						
Last hour	-0.232***	-0.339***	0.031	-0.200***	-0.240***	-0.053
	(0.037)	(0.05)	(0.061)	(0.037)	(0.049)	(0.063)
Second hour	-0.069***	-0.089***	0.168***	-0.067***	-0.064**	0.101***
	(0.019)	(0.025)	(0.035)	(0.019)	(0.025)	(0.036)
Third hour	-0.016	-0.027	0.176***	-0.013	-0.007	0.122***
	(0.015)	(0.020)	(0.028)	(0.015)	(0.020)	(0.029)
Fourth hour	-0.077***	-0.076***	0.087***	-0.052***	-0.044**	0.075***
	(0.014)	(0.017)	(0.026)	(0.014)	(0.017)	(0.027)
Fifth hour	-0.052***	-0.048***	0.046**	-0.037***	-0.029**	0.031
	(0.012)	(0.014)	(0.022)	(0.012)	(0.014)	(0.022)
Sixth hour	-0.032***	-0.028***	0.013	-0.028***	-0.021**	0.002
	(0.008)	(0.010)	(0.014)	(0.008)	(0.010)	(0.015)
Control for time relative to shift beginning	Y	Y	Y	Y	Y	Y
Patient, provider, and other time controls	N	N	N	N	N	N
Sample	Full, actual	$\overline{o} \leq 1$, actual	$\overline{o} \geq 2$, actual	Full, simulated	$\overline{o} \leq 1$, simulated	$\overline{o} \geq 2$, simulated
Number of observations	334,955	231,576	101,657	334,783	231,710	101,692
Adjusted R-squared	0.010	0.011	0.001	0.009	0.011	0.001
Sample mean outcome	-0.920	-0.987	-0.789	-0.927	-0.973	-0.798

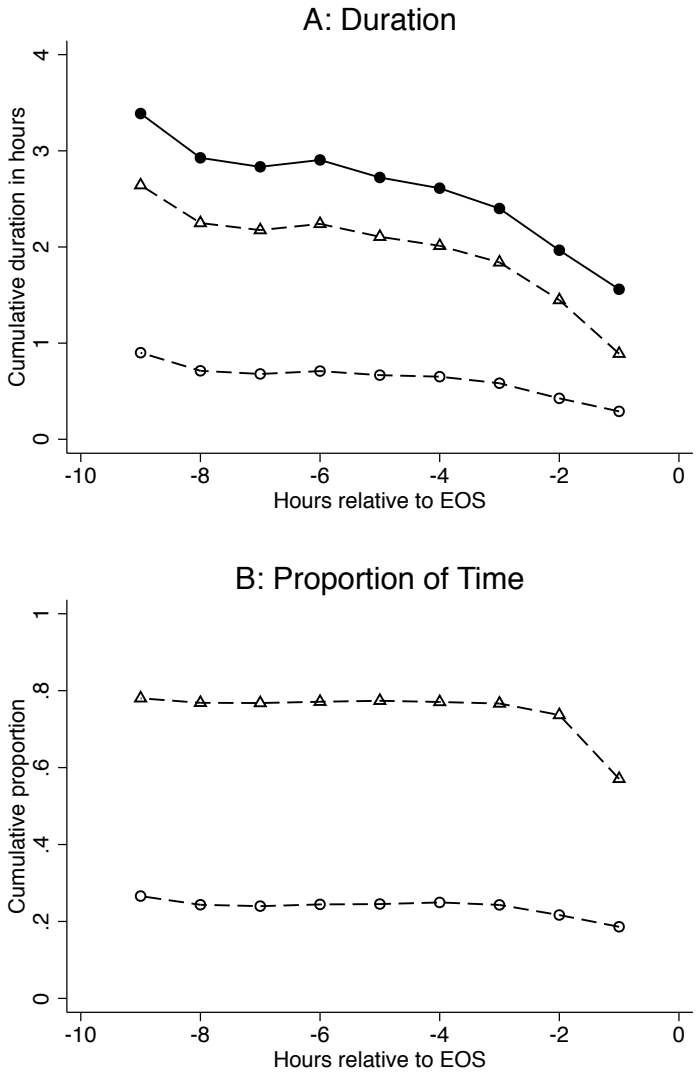
Note: This table is a continuation of Table 3, reporting coefficient estimates and standard errors in parentheses for EOS effects on workload-adjusted length of stay, for arrival at each hour prior to end of shift (EOS), where arrival greater than six hours is the reference period. Models (1) to (3) are estimated on actual data, while Models (4) to (6) are estimated on simulated data. Models also differ by which shifts, based on overlap $\overline{o}$, are included. All models are estimated with Equation (A-4.20), which controls for time relative to shift beginning but not for other variables, in order to facilitate comparison between actual and simulated data. Workload-adjusted length of stay is calculated by Equation (5). * denotes significance at 10% level, ** denotes significance at 5% level, and *** denotes significance at 1% level. Results are graphically shown in Figure A-5.5.

Figure A-1.1: Effects on Length of Stay by Shift Length



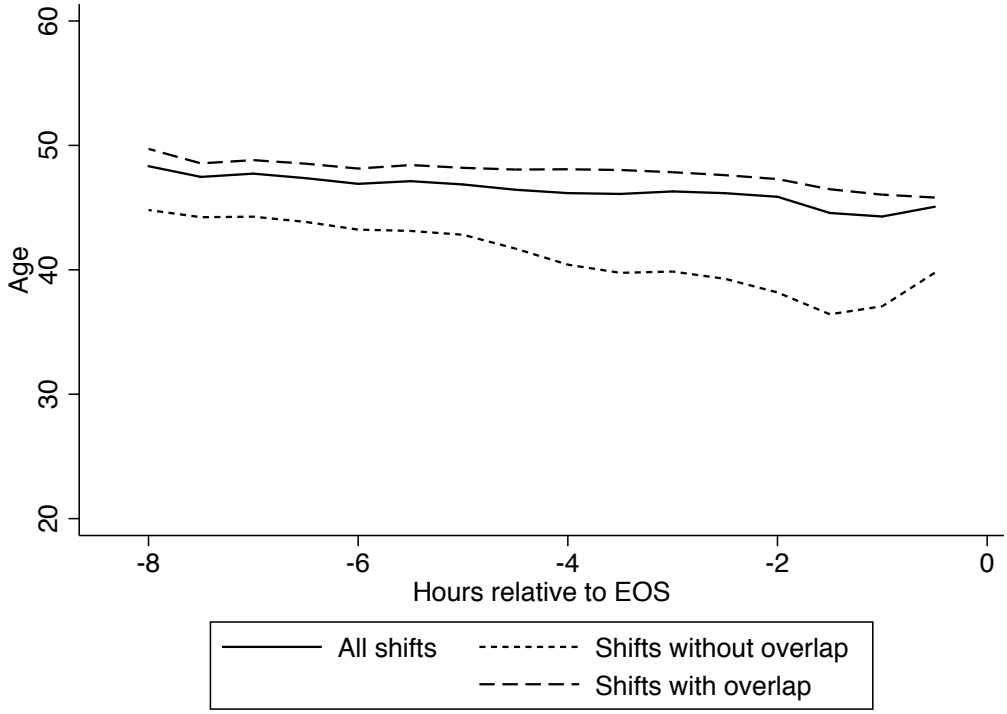
Note: This figure shows coefficients from Equation (2) estimated separately for shifts of seven or eight hours in length (open circles), nine hours in length (closed circles), and ten hours in length (open triangles). Panel A arranges estimates by hours relative to end of shift (EOS). Panel B arranges estimates by hours relative to shift beginning.

Figure A-2.1: Time Components



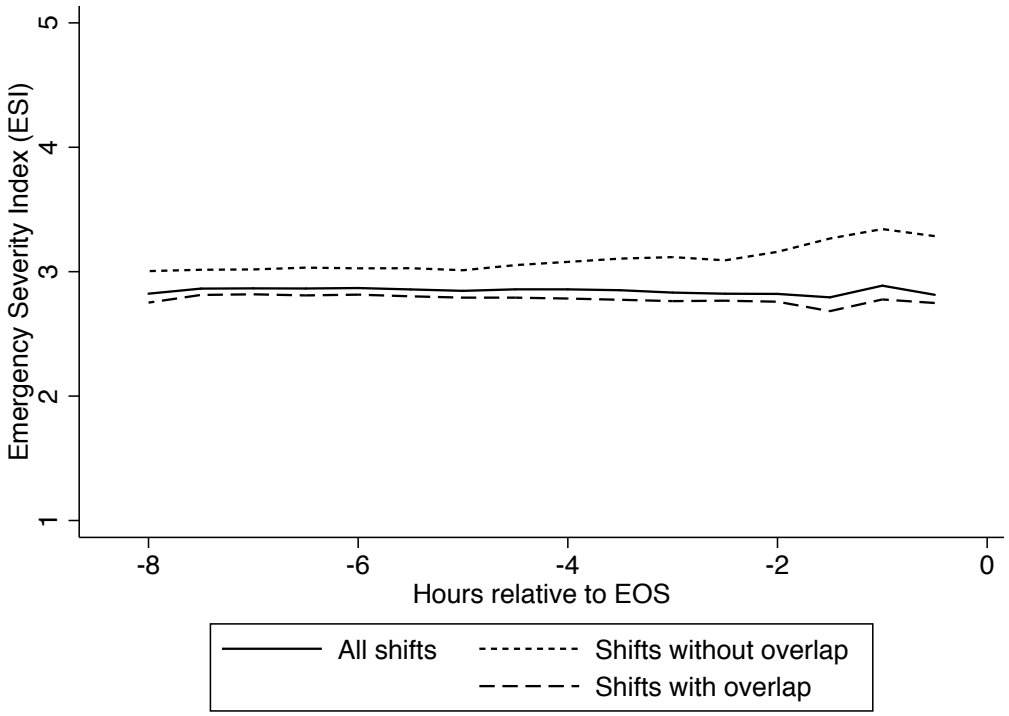
Note: This figure plots time components of length of stay as a function of hours relative to end of shift (EOS): time from pod arrival to first order (open circles), time from first to last (non-discharge) order (open triangles), and time from last order to discharge order (closed circles). Panel B shows marginal effects from a fractional logit model on these shares. Panel A represents these results as time in hours, incorporating results on the EOS effect on length of stay.

Figure A-3.1: Mean Age of Accepted Patients by Arrival Time



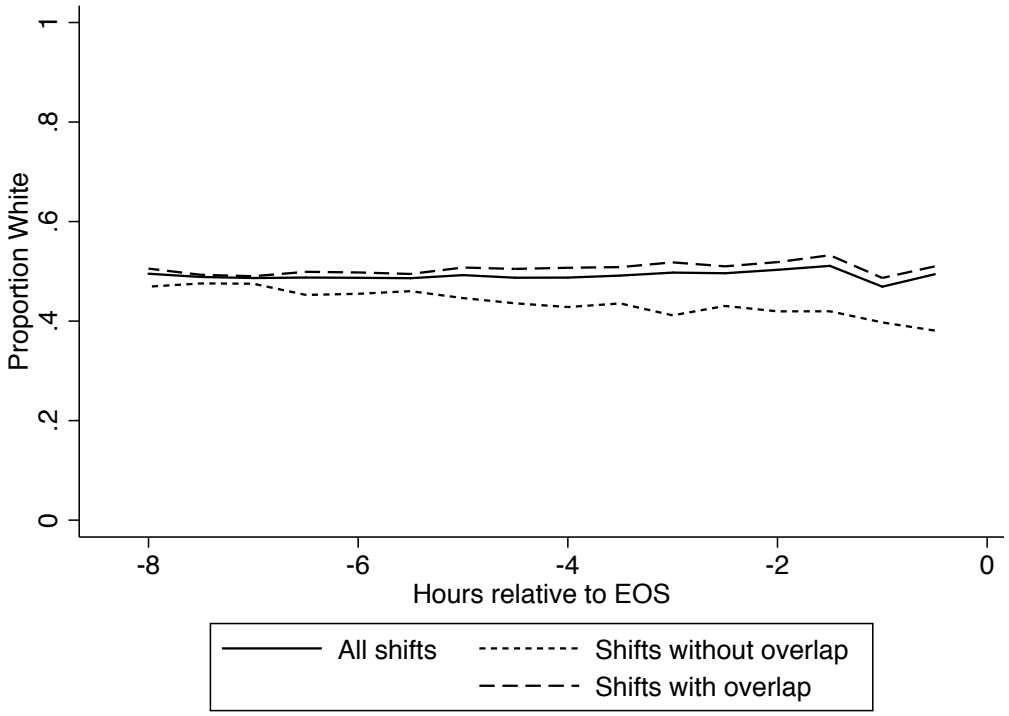
Note: This figure shows mean age of accepted patients in each 30-minute bin of arrival time. The solid line shows means across all shifts. The short-dashed line shows means across shifts without overlap ("terminal shifts"), in which patient "acceptance" near EOS is simply patient assignment by the triage nurse to the managerial location. The long-dashed line shows means across shifts with overlap ("transitioned shifts"), in which patients are chosen between at least two physicians near EOS.

Figure A-3.2: Mean ESI of Accepted Patients by Arrival Time



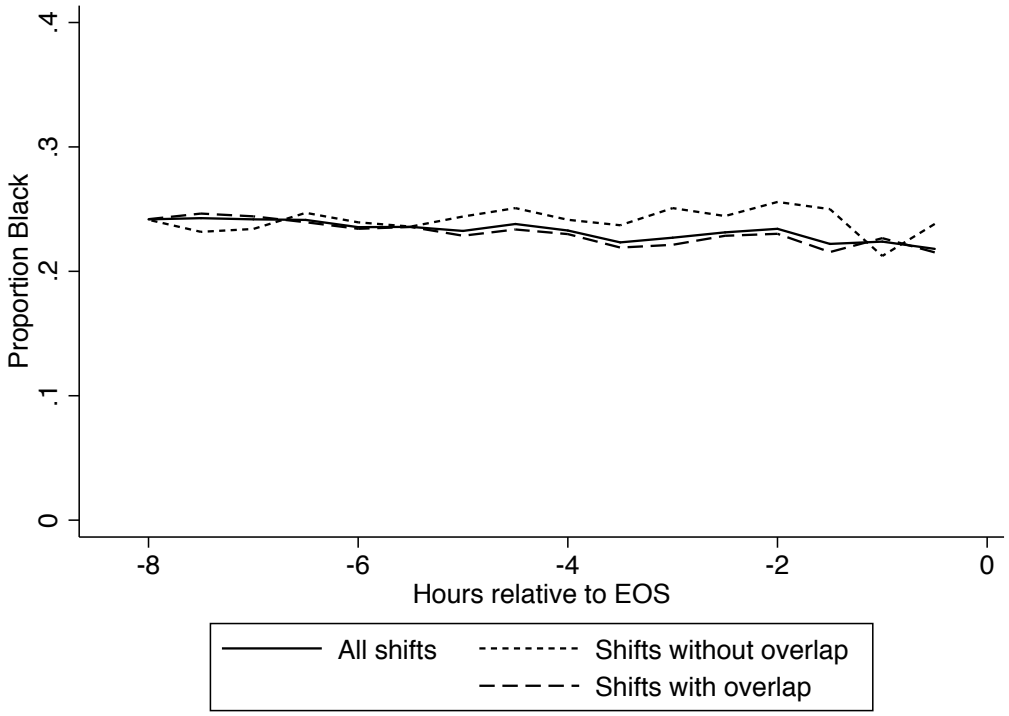
Note: This figure shows mean Emergency Severity Index (ESI) of accepted patients in each 30-minute bin of arrival time. ESI is an integer from 1 (most severe) to 5 (least severe), evaluated by the triage nurse and determined by algorithm (Tanabe et al., 2004). The solid line shows means across all shifts. The short-dashed line shows means across shifts without overlap ("terminal shifts"), in which patient "acceptance" near EOS is simply patient assignment by the triage nurse to the managerial location. The long-dashed line shows means across shifts with overlap ("transitioned shifts"), in which patients are chosen between at least two physicians near EOS.

Figure A-3.3: Proportion White of Accepted Patients by Arrival Time



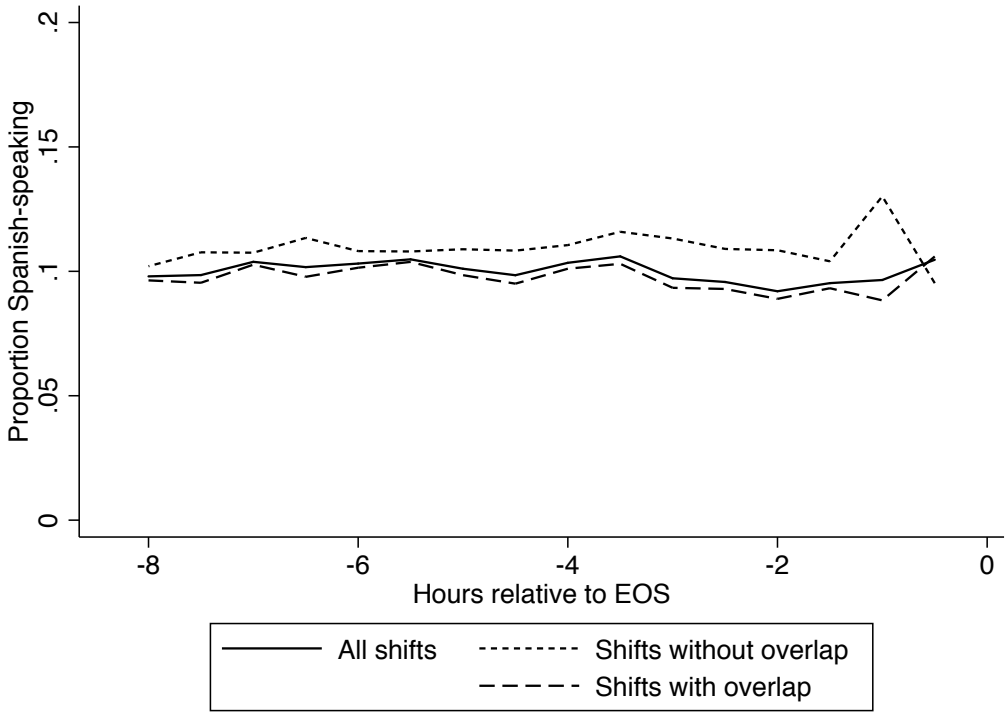
Note: This figure shows the proportion white race of accepted patients in each 30-minute bin of arrival time. The solid line shows proportions across all shifts. The short-dashed line shows proportions across shifts without overlap ("terminal shifts"), in which patient "acceptance" near EOS is simply patient assignment by the triage nurse to the managerial location. The long-dashed line shows proportions across shifts with overlap ("transitioned shifts"), in which patients are chosen between at least two physicians near EOS.

Figure A-3.4: Proportion Black of Accepted Patients by Arrival Time



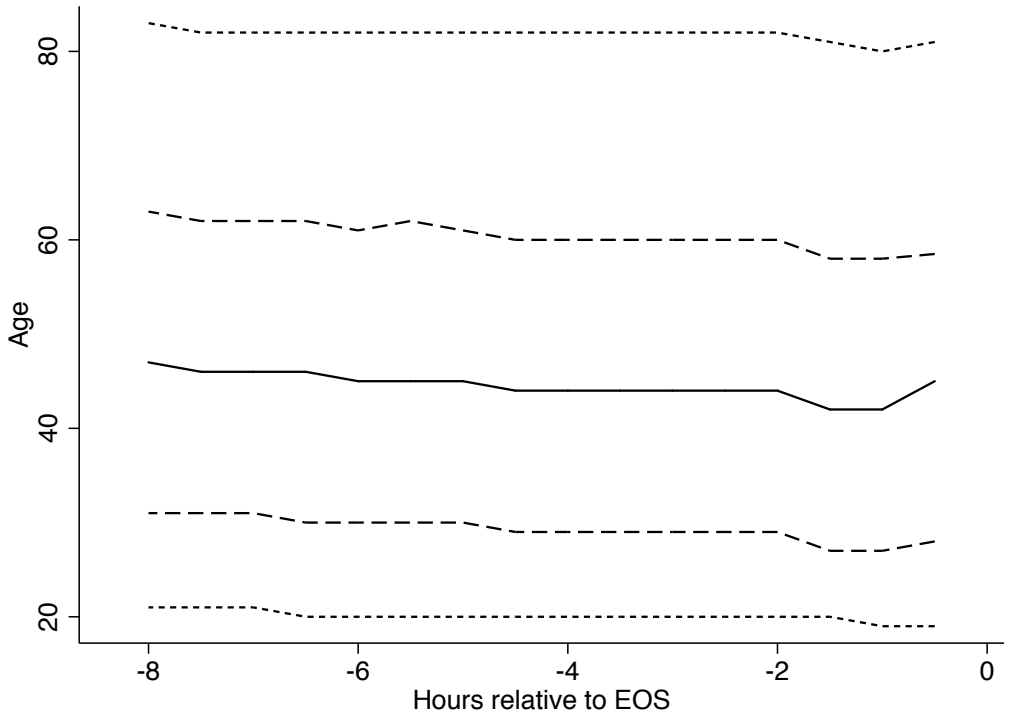
Note: This figure shows the proportion black race of accepted patients in each 30-minute bin of arrival time. The solid line shows proportions across all shifts. The short-dashed line shows proportions across shifts without overlap ("terminal shifts"), in which patient "acceptance" near EOS is simply patient assignment by the triage nurse to the managerial location. The long-dashed line shows proportions across shifts with overlap ("transitioned shifts"), in which patients are chosen between at least two physicians near EOS.

Figure A-3.5: Proportion Spanish-speaking of Accepted Patients by Arrival Time



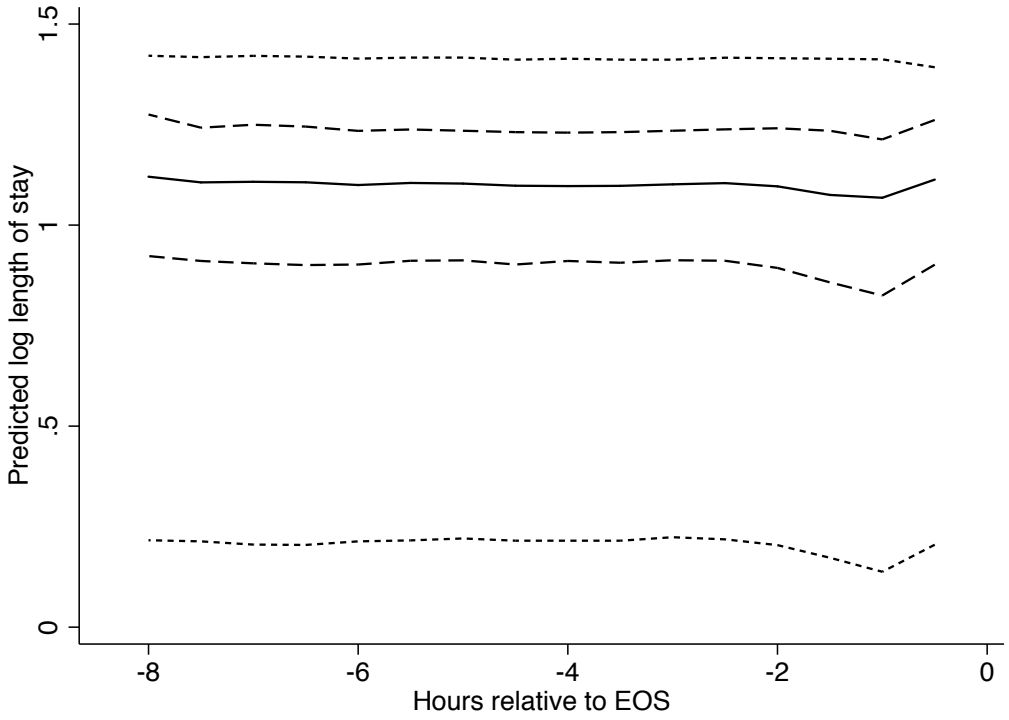
Note: This figure shows the proportion Spanish-speaking of accepted patients in each 30-minute bin of arrival time. The solid line shows proportions across all shifts. The short-dashed line shows proportions across shifts without overlap ("terminal shifts"), in which patient "acceptance" near EOS is simply patient assignment by the triage nurse to the managerial location. The long-dashed line shows proportions across shifts with overlap ("transitioned shifts"), in which patients are chosen between at least two physicians near EOS.

Figure A-3.6: Age Quantiles of Accepted Patients by Arrival Time



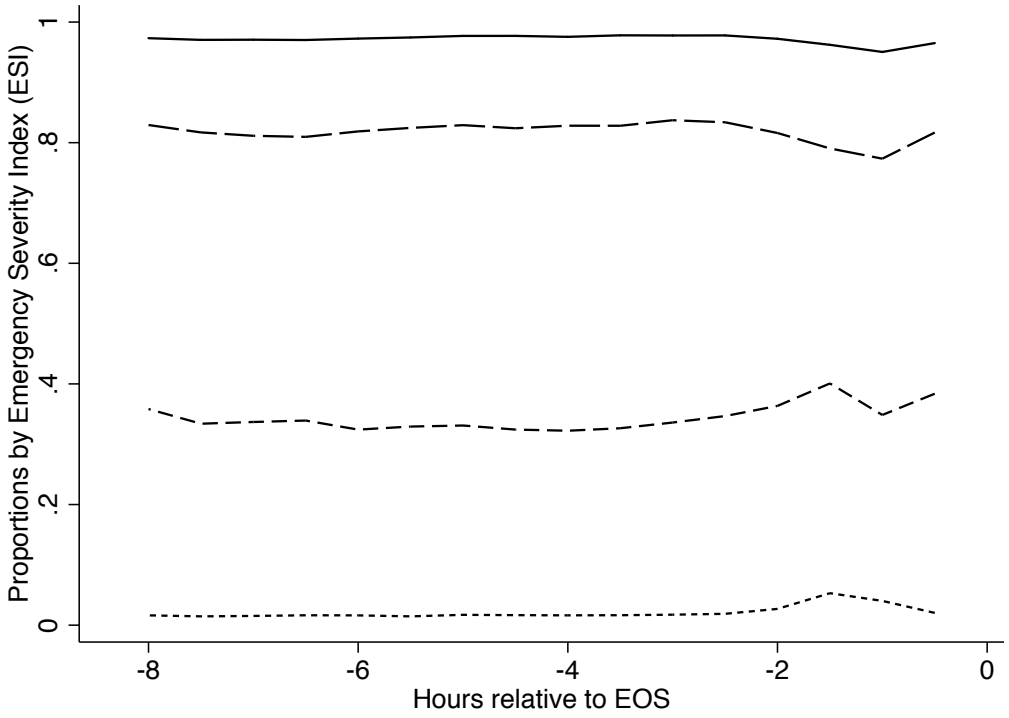
Note: This figure shows age quantiles of accepted patients in each 30-minute bin of arrival time. The solid line shows medians; dashed lines show 25th and 75th percentiles; and short-dashed lines show 5th and 95th percentiles.

Figure A-3.7: Predicted Length of Stay Quantiles of Accepted Patients by Arrival Time



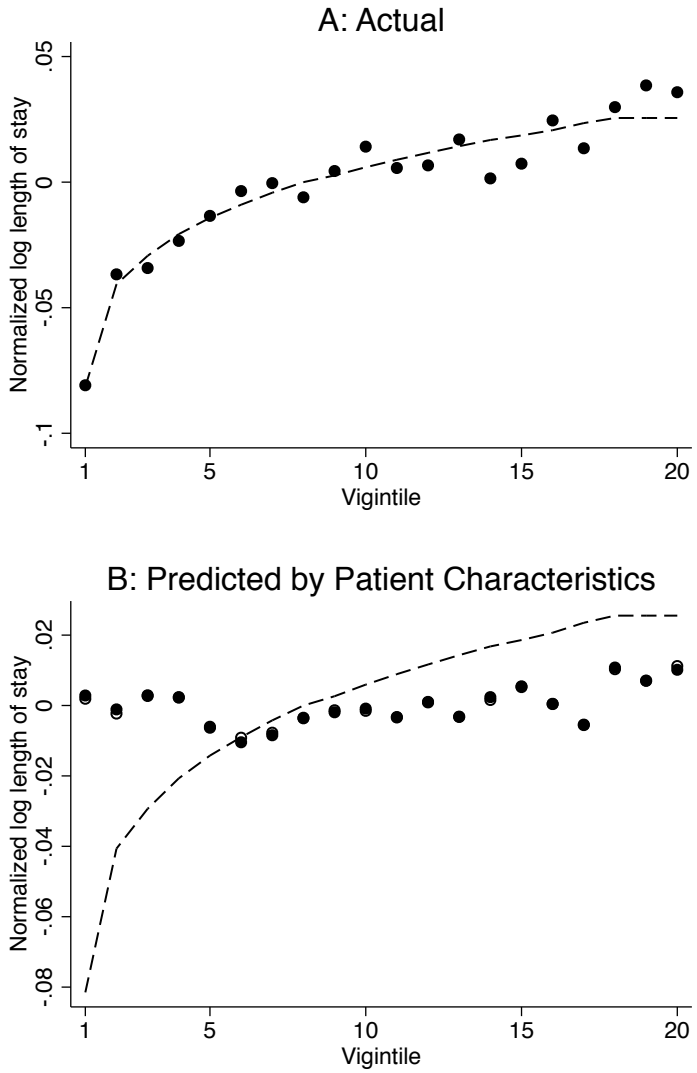
Note: This figure shows predicted log length of stay quantiles of accepted patients in each 30-minute bin of arrival time. Log length of stay is preeicted by cubic splines of age, an indicator for sex, indicators for ESI, indicators for language, and indicators for race. The solid line shows medians; dashed lines show 25th and 75th percentiles; and short-dashed lines show 5th and 95th percentiles.

Figure A-3.8: ESI Proportions of Accepted Patients by Arrival Time



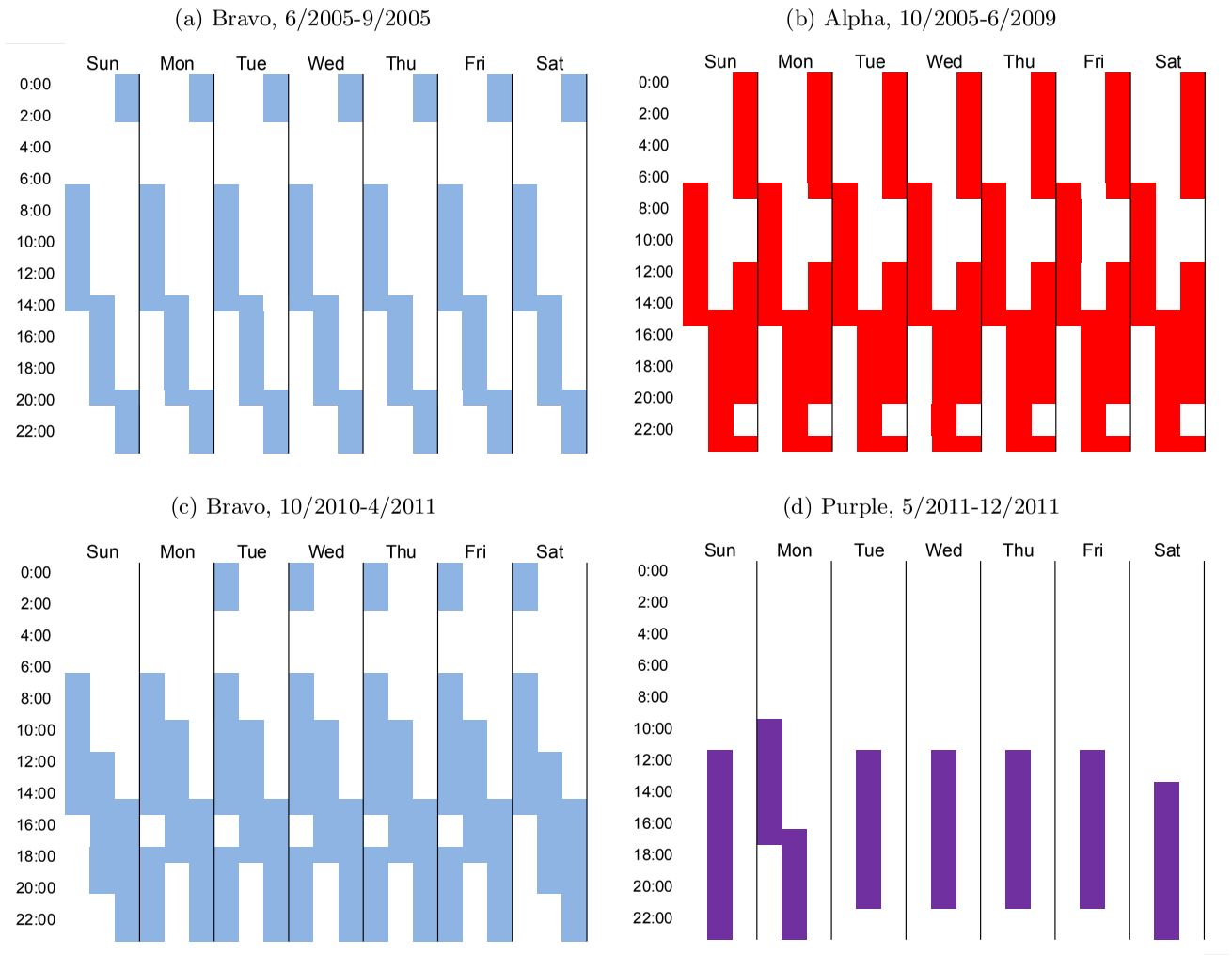
Note: This figure shows (cumulative) proportions by Emergency Severity Index (ESI) of accepted patients in each 30-minute bin of arrival time. The dotted line shows the proportion of patients with ESI 1; the short-dashed and long-dash shows the proportion of patients with ESI at least 2 and 3, respectively; the solid line shows the proportion ofp atients with ESI at least 4.

Figure A-3.9: Actual and Predicted Mean Log Length of Stay



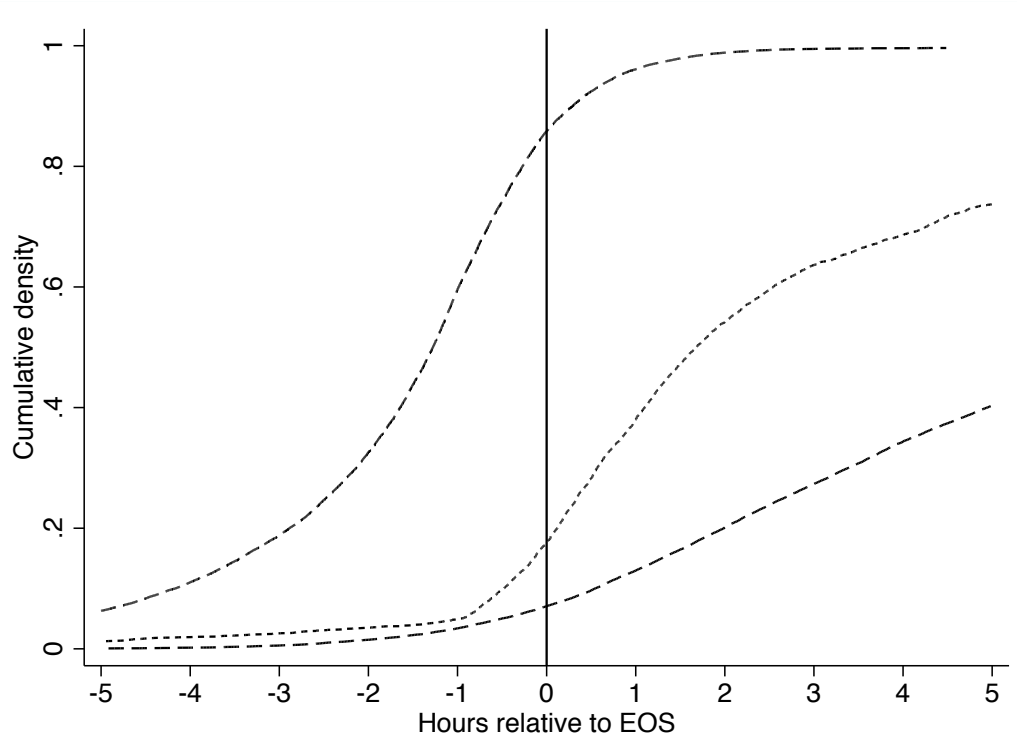
Note: This figure presents binned results in Figure 6 in a different manner. The dashed line in both panels represents mean hourly predictions based on times relative to EOS among shifts for each arrival hour, Q_t. Panel A shows the relationship between Q_t and mean actual log length of stay. Panel B shows the relationship betwen Q_t and predicted log length of stay based on patient characteristics; predictions based on "ex ante" and full patient characteristics are shown as solid and hollow dots, respectively. See note for 6 for more detail.

Figure A-5.1: Example Weekly Pod Schedules



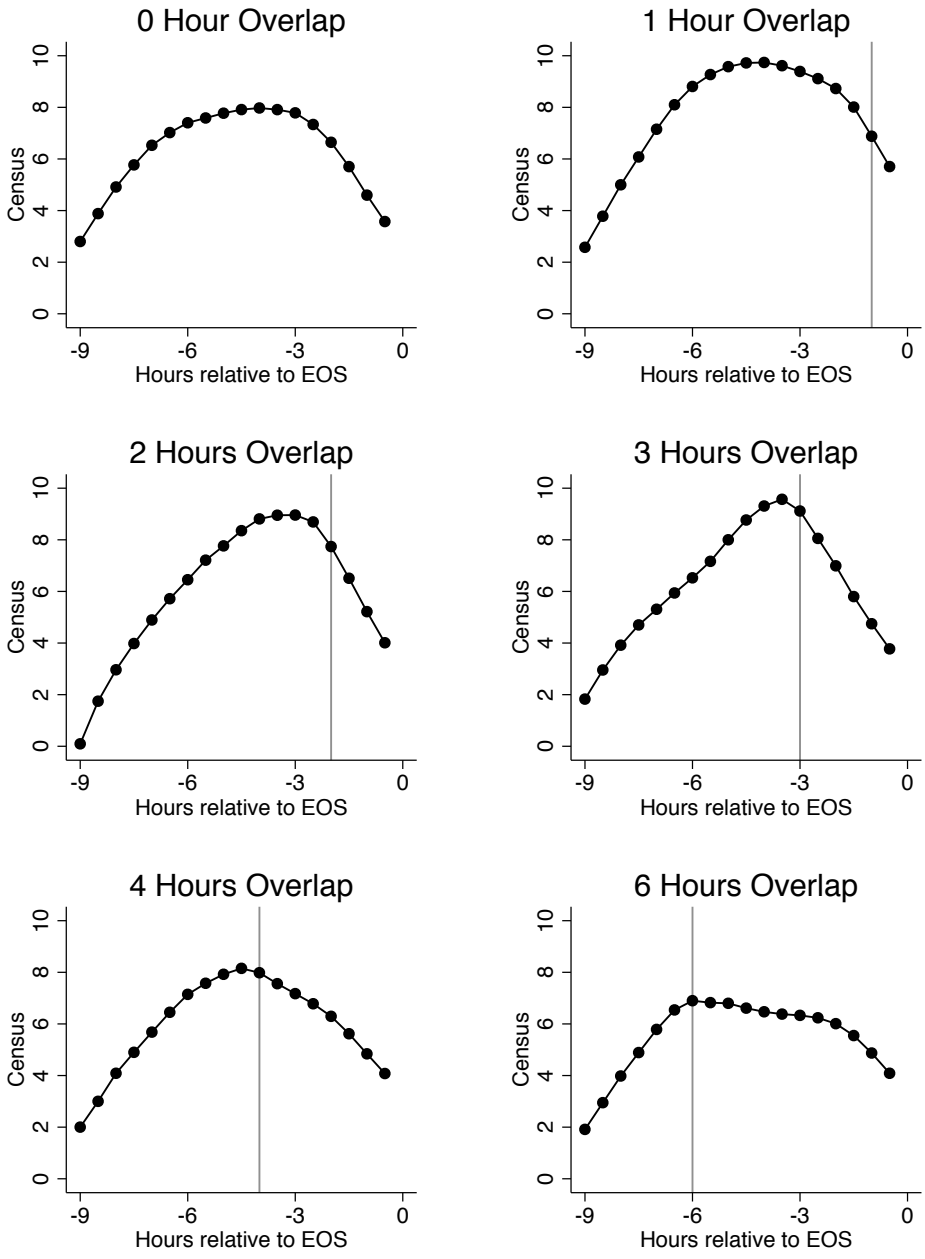
Note: Filled areas in vertical lines represent hours scheduled for a shift for a single physician. Hours when there is more than one physician present are represented by horizontally adjacent filled areas.

Figure A-5.2: Evidence from Orders of Staying Past EOS



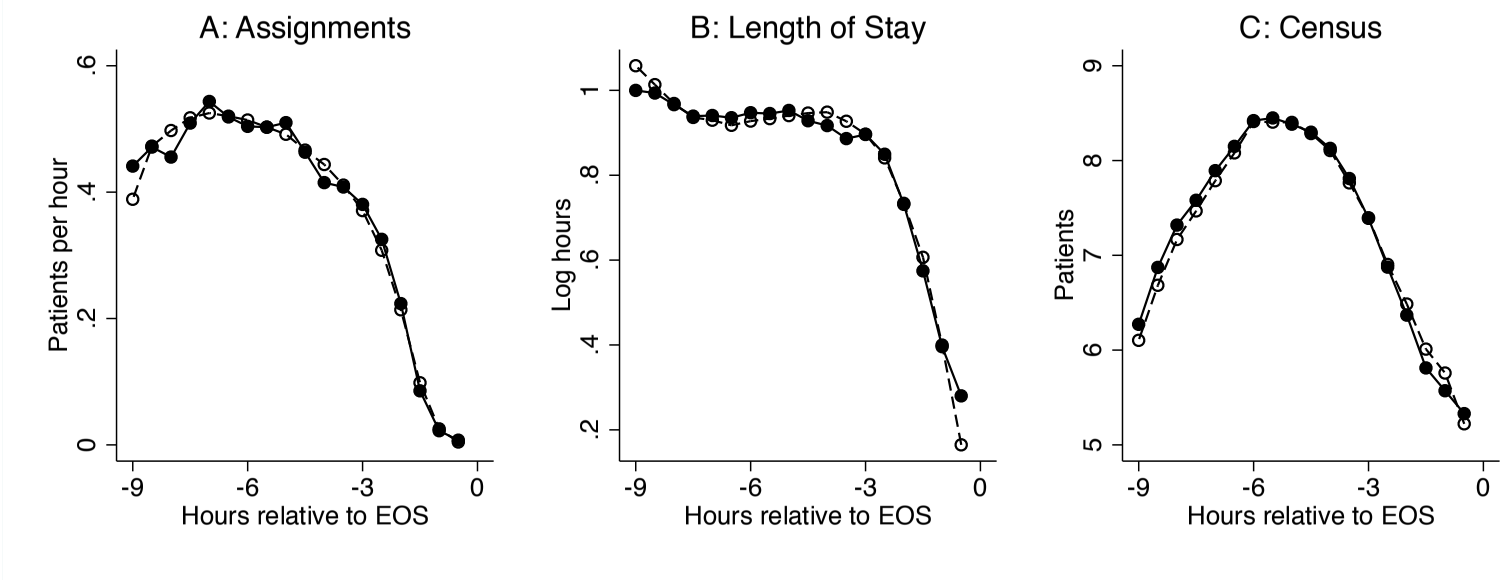
Note: This figure shows the following cumulative density plots, in order from left to right: last order by the attending physician of record (the physician on the bill for patient care, or AOR, corresponding to the physician whose shift is matched to a patient visit), the first (non-resident) physician order after AOR orders, and the last discharge order. The cumulative density of the last order by the AOR at EOS is approximately 86% (i.e., 14% of AORs write orders after their EOS). The median time for the first physician order after AOR orders is 1.7 hours. Only 35% of patient visits have an order by the AOR. Of these visits, only 6% of visits have a subsequent order by an attending physician. Cumulative densities for physician orders are calculated using visits in which the relevant physician order is observed at least once. See notes for this figure in Appendix A-5 for further comments.

Figure A-5.3: Censuses over Time



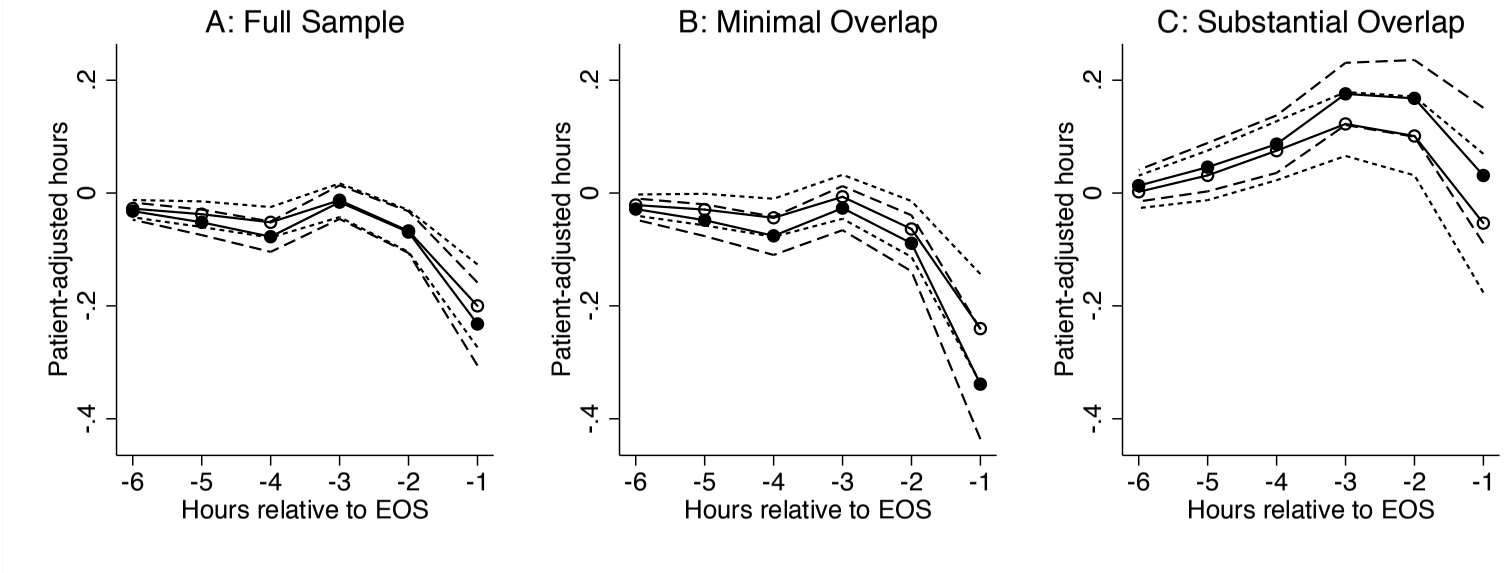
Note: This figure plots average censuses over time relative to the end of shift (EOS). Each panel shows results for physicians in shifts with a given EOS overlap time. Subsequent shift starting times are marked with a vertical line.

Figure A-5.4: Model Fit



Note: This figure shows the fit of the structural model described in Section 7 in terms of patient assignments (Panel A), length of stay (Panel B), and census (Panel C), averaged in each 30-minute interval relative to end of shift (EOS). Average actual data are shown in solid circles; average simulated data are shown in hollow circles.

Figure A-5.5: Workload-adjusted Length of Stay



Note: This figure shows coefficients for regressions, described in Equation (A-4.20), of the log of workload-adjusted length of stay (length of stay divided by average census) on time relative to end of shift (EOS), controlling for time from beginning of shift using both actual data (closed circles, confidence intervals in long-dashed lines) and simulated data (open circles, confidence intervals in short-dashed lines). Panel A shows results using actual or simulated data for all shifts. Results using actual or simulated data either only for shifts where EOS overlap $\overline{o} \leq 1$ or only for shifts with $\overline{o} \geq 2$ are shown in Panels B and C, respectively. Numbers for this figure are shown in Table A-5.3.