ECONOMETRICA

JOURNAL OF THE ECONOMETRIC SOCIETY

An International Society for the Advancement of Economic Theory in its Relation to Statistics and Mathematics

http://www.econometricsociety.org/

Econometrica, Vol. 86, No. 3 (May, 2018), 997–1030

THE EFFICIENCY OF SLACKING OFF: EVIDENCE FROM THE EMERGENCY DEPARTMENT

DAVID C. CHAN
Stanford University

Econometrica, Vol. 86, No. 3 (May, 2018), 997–1030

THE EFFICIENCY OF SLACKING OFF: EVIDENCE FROM THE EMERGENCY DEPARTMENT

DAVID C. CHAN
Stanford University

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 assigned 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 more than six times their market wage to preserve their leisure. Accounting for overall costs, I find that physicians slack off at approximately second-best optimal levels.

KEYWORDS: Scheduling, shift work, optimal work assignment, physicians.

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 specifying compensation based on availability (or *minimum quantity* of hours worked), schedules may open a margin for distortionary behavior if the appropriate time to complete work tasks is private information.

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 assigned tasks near EOS. These effects are particularly important in work that is uncertain, time-sensitive, and information-rich.[2] For example, if tasks are un-

David C. Chan: david.c.chan@stanford.edu

I am thankful to Jason Abaluck, Panle Barwick, Lanier Benkard, Amitabh Chandra, Daniel Chen, David Cutler, Mike Dickstein, Joe Doyle, Liran Einav, Bob Gibbons, Jeremy Goldhaber-Fiebert, Jon Gruber, Igal Hendel, Rob Huckman, Dan Kessler, Eddie Lazear, Robin Lee, Sara Machado, Grant Miller, Paul Oyer, Maria Polyakova, Stephen Ryan, Orie Shelef, Matt Shum, 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 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.

[1]A large and active literature in operations management has viewed scheduling workers as mechanical inputs (e.g., Perdikaki, Kesavan, and Swaminathan (2012), He, Dexter, Macario, and Zenios (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, for example, 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

 https://doi.org/10.3982/ECTA13565

certain and time-sensitive, as they are in many environments with scheduling, then they cannot be pre-assigned. Performance on information-based tasks is often difficult to evaluate and therefore non-contractible. Finally, once a worker is assigned a task, worker-task specificity, for example due to tacit knowledge or personal relationships intrinsic to the task, makes passing the task to another worker costly.

In this paper, I study the effects of schedules on the behavior of emergency department (ED) physicians working in shifts. Shifts ending at different times, particularly due to changes in the shift schedule, allow me to separate effects related to shift work from differences due to the time of day. Shifts of different lengths allow separating 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 are assigned, 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 patient arrival approaches EOS. To support the claim that I have identified causal effects of time to EOS on patient care, I utilize patient characteristics plausibly unobserved at the time of assignment and quasi-random variation in the propensity of assignment to a shift nearing EOS as a function of patient arrival times at the ED.

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 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 efficient patient-care decisions. In other words, conditional on these other characteristics of the work environment, the EOS is merely an arbitrarily varying rule stipulating 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. Further, distortions appear concentrated in shifts ending during daylight hours, when non-work plans may be costlier to defer. When EOS is during daylight, physicians are less likely to write orders past EOS, and they shorten patient stays to a greater extent when approaching EOS, consistent with physicians seeking leisure.

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. Although assigning more patients to physicians near EOS would reduce slacking off, it would increase the workload near EOS and the pressure for physicians to rush, inefficiently substituting other inputs for time.[3] I analyze this tradeoff in a stylized theoretical model featuring schedules, worker-task specificity, and non-contractible performance of tasks. This model show that workers prefer fewer tasks (i.e., "slack off") relative to the first-best assignment near EOS, but when performance on the tasks cannot be contracted upon, the (second-best) optimal assignment policy still allows some slacking off.

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 portion of the economy with irregular or non-standard work times has grown significantly (Beers (2000), Katz and Krueger (2016)). Key questions in these settings are how to schedule worker availability and how to assign tasks to workers.

[3]The idea of time per effective work is related to work by Coviello, Ichino, and Persico (2014), who discussed 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 assess the observed patient-assignment policy relative to counterfactual policies along the dimension of slacking off, I specify and estimate a discrete choice dynamic programming model. In this model, physicians care about how they discharge individual patients and the number of patients they are left with at EOS. As physicians proceed through their shifts, their choices to discharge patients are influenced by expectations of future patient assignments. The model allows me to simulate costs of physician time, patient time, and hospital resources under counterfactual assignment policies. Assigning more patients near EOS such that physicians stay an additional hour induces an additional $3300 in hospital spending per shift; physicians also reveal that they are willing to spend more than $790 in hospital dollars per each hour of leisure saved, which is more than six times greater than the market wage. Consistent with the high marginal rate of technical substitution between physician time and hospital resources, overall costs rise steeply under counterfactual policies that assign more patients near EOS but change relatively little under policies that assign fewer patients. Interestingly, I find that the observed assignment policy approximately minimizes overall costs.

This paper contributes to two strands of literature. First, a central economic question is how 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] As work environments featuring flexible or irregular hours grow in prominence, recent empirical work has examined the effects of such workplaces on productivity (Bloom, Liang, Roberts, and Ying (2015)) and workers' underlying preferences for such arrangements (Moen, Kelly, Fan, Lee, Almeida, Kossek, and Buxton (2016), Mas and Pallais (2016)).

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.

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 dramatically increased the number of diagnostic and therapeutic decisions

[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 (2017), Larkin (2014)).

[5]The general case of intrinsic motivation has been discussed by Tirole (1986) and in later papers (Dewatripont, Jewitt, and Tirole (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, for example, peer effects due to social incentives (Bandiera, Barankay, and Rasul (2005, 2009), Mas and Moretti (2009)) and the response to information arguably orthogonal to profits (Kolstad (2013)).

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 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, Devers, Lake, Reed, and Stoddard (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 investigates patient assignment rates relative to EOS. Section 4 reports EOS effects for patients who are assigned. Section 5 considers the relationship between shift overlap, workload, and patient-care distortion. Section 6 analyzes a stylized theoretical model of the optimality of slacking off. Section 7 presents results from a dynamic programming model to consider counterfactual policies 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 in the United States 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-7.1).

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 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 (see Figure 1). 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. 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

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

[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 $\overline{o} > 0$.

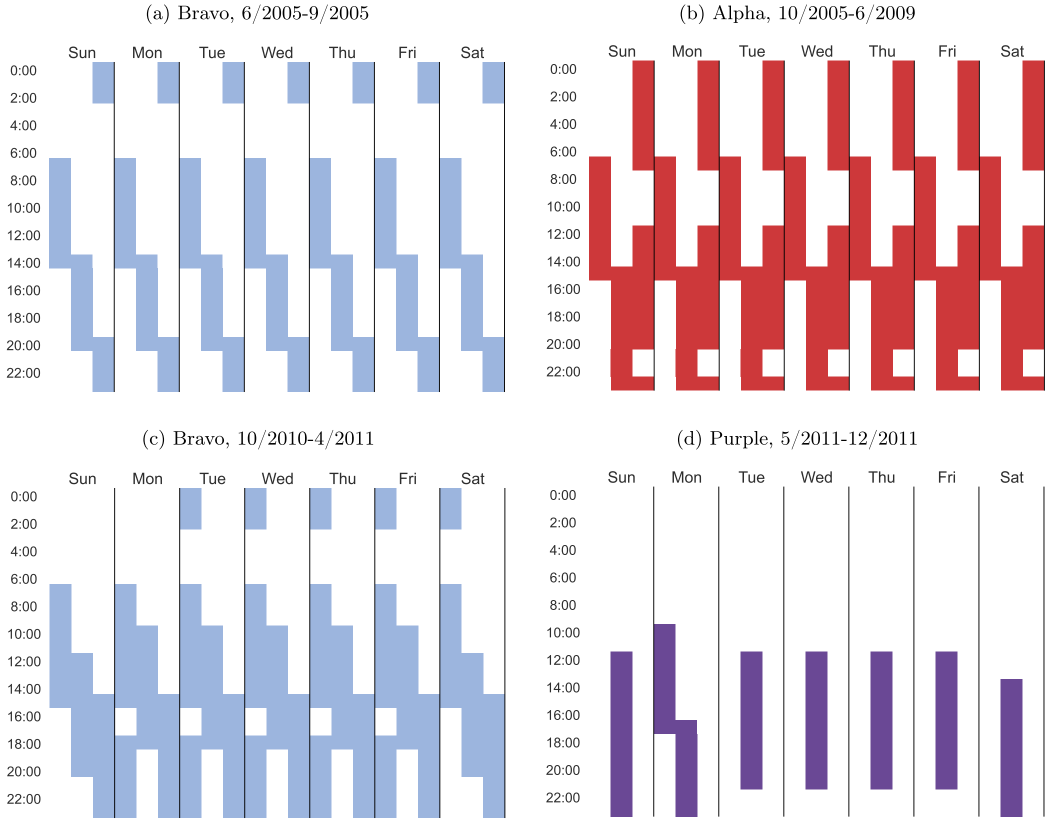

FIGURE 1.—Example weekly pod schedules. *Note*: Filled areas in vertical lines represent hours scheduled for a shift for a single physician. Hours when there is more than one physician present are represented by horizontally adjacent filled areas.

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% bonus based on clinical revenue (measured by Relative Value Units, or RVUs, per hour) and modified by research, teaching, and administrative metrics.[8] Although their salaries are based on numbers of shifts worked, physicians are not compensated for time worked past EOS.[9]

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 physi-

[8]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.

[9]This is the standard financial arrangement for salaried physicians across the United States. 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).

cians who may assume care for the patient. Once the patient arrives in a bed, a physician may sign up for that patient, if the patient is in her managerial location. 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, Mallak, and Gibson (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.[10] 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 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.[11]

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 to bronchodilators for suspected asthma. But time—for a careful history-taking, physical examination, 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

[10]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 attending physician orders is presented in Figure A-7.1.

[11]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.

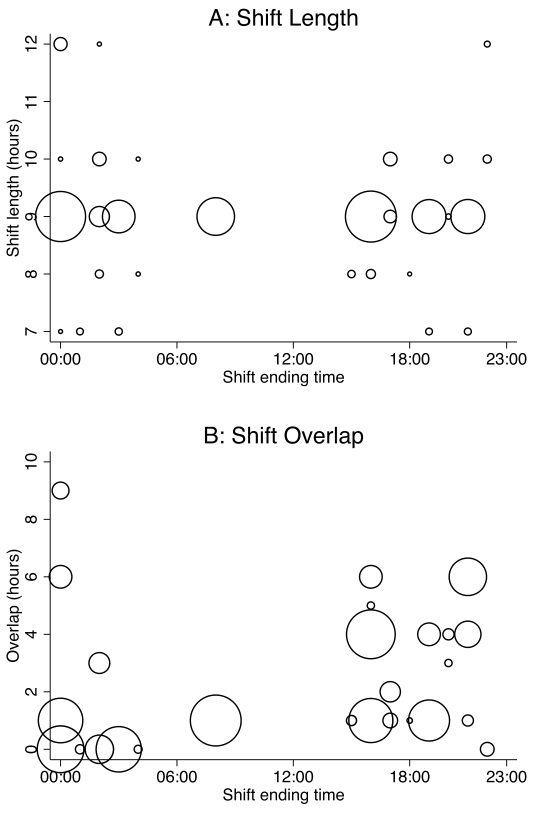

FIGURE 2.—Shift variation. *Note*: This figure illustrates the variation in observations across shift types. Panel A plots shifts by shift ending time and shift length. Panel B plots shifts by shift ending time and the length of overlapping transition at the end of shift. The sizes of circles in both panels represent the number of shifts out of 23,990 total in each two-dimensional category.

of the sample definition process are described in Table A-7.2. In the sample, I observe the identities of 102 physicians, 1146 residents and physician assistants, and 393 nurses.

Because I focus on behavior near EOS, I present in Figure 2 the key variation across the 23,990 shifts in the time of day for EOS, shift length, and the overlap with another shift at EOS. Table A-7.1 lists the underlying 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.

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 Appendix A-3.

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 also received attention as a source of rising system costs (Schuur and Venkatesh (2012), Forster, Stiell, Wells, Lee, and Van Walraven (2003)). I accordingly measure total direct costs, which are the hospital's internal measures of costs incurred both in the ED and possibly during a subsequent admission.[12] Finally, I measure 30-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, Gimbel, Yarnold, Kyriacou, and Adams (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).

In addition to elements displayed via the ED computer interface, I observe patient language, race, zip code of residence, and diagnostic information. The last characteristic of diagnoses is only incompletely known by physicians (or anyone) prior to assignment (via the "chief complaint"), especially since physicians do not interact with patients or examine their medical records 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, Steiner, Harris, and Coffey (1998)). Diagnostic codes are recorded after patients are seen by the physician and discharged and therefore are also partly determined by patient care.

3. PATIENT ASSIGNMENT

In this section, I describe patient assignment near EOS, a function of both triage nurse choices to assign patients to locations and physician choices to "accept" patients in their location. Figure 3 presents the hourly average rates of new patient assignments in 30-minute bins relative to EOS. Each panel represents shifts with a different EOS overlap, $\overline{o}$, and shows assignments for the index physician (patients accepted) and for the location inclusive of the index physician (patients assigned by the triage nurse). Physicians are generally assigned two or three new patients in an hour, and assignment rates are highest near the beginning of shift. For $\overline{o} > 0$, assignment rates show two relationships with time. First, patient flow declines 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. For $\overline{o} = 0$, 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 assignment to the leaving physician, patients continue to arrive at the pod at similar or greater rates prior to the peer's transitioning shift.

[12]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). These costs are internal valuations of actual resources. They are not charges and are unrelated to billing or revenue source.

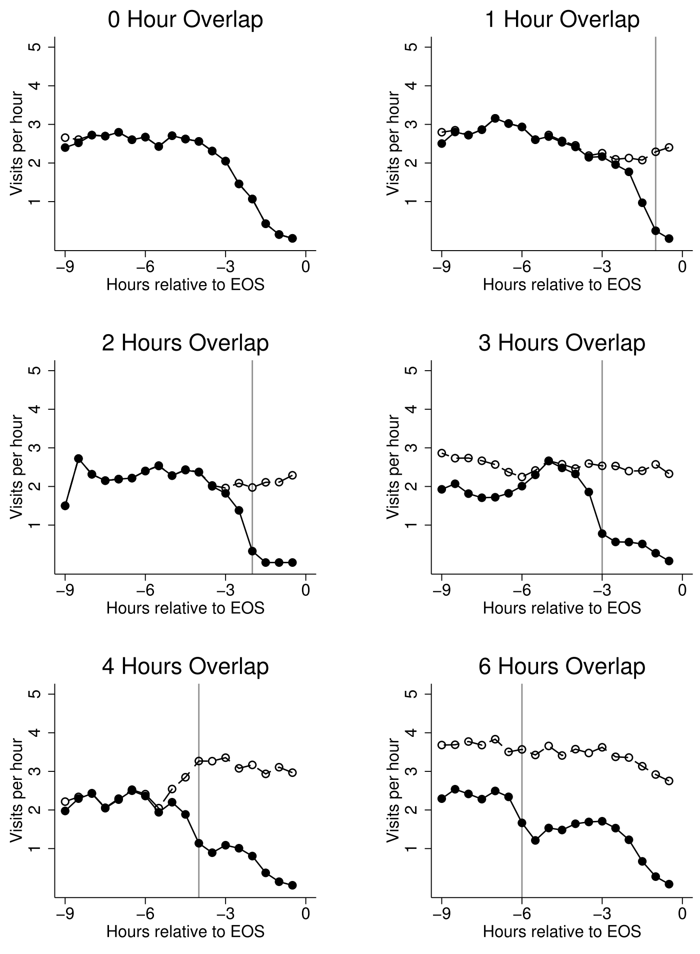

FIGURE 3.—Flow of patient visits over time. *Note*: This figure shows unadjusted average hourly rates of patient visits for each 30-minute interval relative to end of shift (EOS). Each panel shows results for shifts with a given EOS overlap time. Patient visits for the index physician are shown in closed circles; patient visits for the location are shown in open circles. Subsequent shift starting times are marked with a vertical line.

The earlier arrival of peers allows for earlier reductions in patient assignment relative to EOS. The reductions are in fact *prior* to peer arrival, especially in shifts with shorter transitions, suggesting forward-looking behavior. For terminal shifts with no other physician working near EOS in the same location, the long decline in patient flow rates results from the *triage nurse* assigning fewer patients to the location. Thus, assignment to physicians in general and "slacking off" in particular is achieved both by coworkers sharing a location and by triage nurse assignment to locations.

4. EFFECT ON PATIENT CARE

4.1. *Identification and Balance*

The panel nature of the data allows for me to control for time categories (e.g., time of the day or day of the week) because shifts start at different times. Furthermore, variation in shift lengths identifies the EOS effect separately from fatigue or other effects that

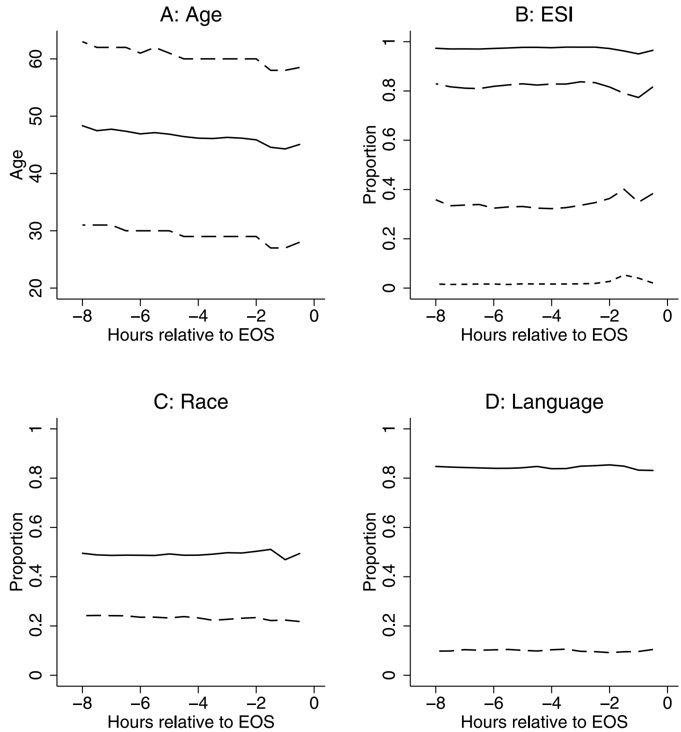

FIGURE 4.—Patient characteristics over time. *Note*: This figure shows statistics of the distribution of characteristics for patients assigned to physicians in each 30-minute bin relative to EOS. Panel A shows mean age (solid line) and 25th and 75th percentiles of age (dashed lines). Panel B shows cumulative proportions of patients with an emergency severity index (ESI) from 1, ≤ 2, ≤ 3, and ≤ 4, respectively. 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)). Panel C shows proportions of patient white (solid) and black (dashed) race. Panel D shows proportion of English (solid) and Spanish (dashed) language.

depend on time relative to the beginning of shifts.[13] Finally, I observe the same physicians and support staff (i.e., physician assistants, nurses, and residents) and can therefore control for their identities as a given physician approaches EOS. However, given that the number of assigned patients clearly declines as physicians near EOS, an important question is how the EOS effect on patient care is identified separately from patient selection correlated with time to EOS.

Figure 4 shows statistics of characteristics for patients assigned to physicians in 30-minute bins relative to EOS. Compared to numbers of patients assigned (Figure 3), and compared to variation *within* bins, patient characteristics are relatively constant across bins. There is a slight trend toward younger patients being assigned to physicians nearing EOS. Figure 5 similarly shows that the distribution of predicted length of stay, using ex ante characteristics of age, sex, ESI, race, and language that are presumably observable at the time of assignment, is quite stable across time relative to EOS.[14] This descriptive

[13]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.

[14]Appendix A-1.1 in the Supplemental Material (Chan (2018)) quantifies this selection in terms of predicted outcomes.

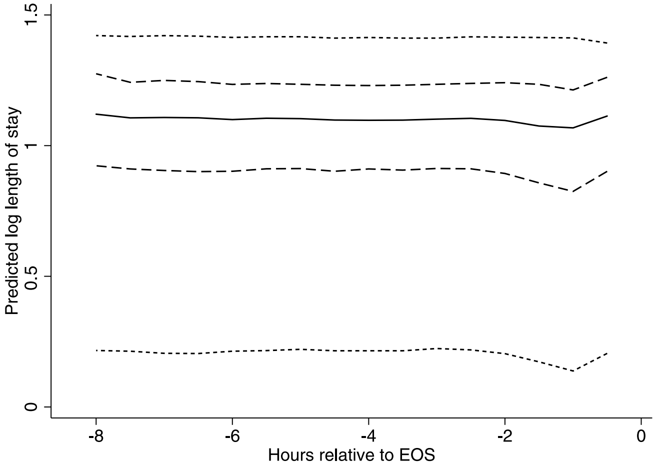

FIGURE 5.—Predicted length of stay over time. *Note*: This figure shows quantiles of predicted log length of stay for patients assigned to physicians in each 30-minute bin relative to EOS. Log length of stay is predicted by cubic splines of age, sex, indicators for ESI, indicators for language, and indicators for race. The solid line shows medians; dashed lines show 25th and 75th percentiles; and short-dashed lines show 5th and 95th percentiles.

evidence suggests that assignment policies relative to EOS are primarily based on patient numbers rather than patient characteristics.

To be clear about identification, I consider two independent assumptions:

ASSUMPTION 1—Excludable Characteristics: *Conditional on ex ante patient characteristics, time of arrival, pod, and providers, patient potential outcomes are mean independent of assigned time relative to EOS.*

ASSUMPTION 2—Random Arrival Times: *Conditional on time categories of arrival (e.g., day of week, time of day), pod, and providers, patient potential outcomes are mean independent of ED arrival times with different propensities for assignment to times relative to EOS.*

The intuition behind Assumption 1 is that assignment within the ED operates through patient numbers and (to a much lesser extent) on ex ante patient characteristics.[15] Other patient characteristics that are correlated with potential outcomes are mostly unknown to physicians before assignment and, under this assumption, excluded from assignment policies. This assumption is testable to the extent that I observe ex post clinical characteristics and can show that, conditional on ex ante patient characteristics, time categories, and staff identities, there exists no correlation between clinical characteristics and assignment to times relative to EOS. The intuition behind Assumption 2 is that, although ED staff may influence patient assignment within the ED, patients arrive at the ED without any systematic selection toward times when there may or may not be a physician near EOS. Specifically, variation in shift schedules within a time category of ED arrival drives

[15]This assumption is strengthened by the institutional fact that triage in the ED is supposed to be sufficiently summarized by the ex ante characteristic of ESI (Tanabe et al. (2004)). Physicians are discouraged from further assessing patients prior to accepting them.

the propensity of being assigned to a physician near EOS but is mean independent of potential outcomes of the arriving patients.[16]

I assess the plausibility of each assumption in a regression framework. To assess Assumption 1, I regress presumably excludable patient characteristics based on Elixhauser diagnoses—specifically (i) predicted length of stay by diagnoses and (ii) count of diagnoses—on hour relative to EOS, controlling for ex ante patient characteristics, time categories (hour relative to beginning of shift, hour of the day, day of the week, and month-year interactions), pod, and provider identities. To evaluate Assumption 2, I regress both ex ante and ex post patient characteristics on hourly propensities for assignment to a physician at a time relative to her EOS, controlling for time categories, pod, and provider identities. I consider (i) predicted length of stay by all patient characteristics, (ii) age, (iii) male sex, (iv) ESI, (v) white race, (vi) black race, (vii) English language, (viii) Spanish language, and (ix) count of Elixhauser indices. Appendix A-1.2 provides further details.

Table I reports results consistent with conditional random assignment under either Assumption 1 (the first two columns) or Assumption 1 (the remaining columns). None of the regressions yield jointly significant coefficients on hour relative to EOS. Furthermore, Table I shows that variation in any of the patient characteristics, residualized by covariates depending on the assumption, remains quite large. In fact, if Assumption 1 were violated, and if patients with the lowest predicted length of stay based on ex post clinical characteristics were perfectly sorted to hours closest to EOS when more than one physician is available, the bias in the last hour of shift would be −0.177, more than 17 standard errors apart from the corresponding coefficient in the corresponding balance test (Column 1 of Table I). Table A-1.1 presents additional balance results for other predicted outcomes, assessing both Assumptions 1 and 2.

4.2. *Main EOS Effects*

In the full specification, I estimate the following equation:

$$Y_{it} = \alpha_{\overline{m}(i,t)} + \gamma_{\underline{m}(i,t)} + \beta \mathbf{X}_i + \eta \mathbf{T}_t + \zeta_{p(i)} + \nu_{j(i),k(i)} + \varepsilon_{it}. \tag{1}$$

Outcome Y_{it} is indexed for patient visit i at time t, and the object of interest is arrival hour $\overline{m}(i,t) = \lceil \overline{t}(i) - t \rceil$ prior to EOS, where seven or greater hours prior to EOS is the omitted category. I control for time relative to the shift beginning $\underline{m}(i,t) = \lfloor t - \underline{t}(i) \rfloor$, patient characteristics $\mathbf{X}_i$, time categories $\mathbf{T}_t$ (for month-year, day of the week, and hour of the day), pod $p(i)$, and physician $j(i)$ and team $k(i)$ (i.e., resident or physician assistant, and nurse) identities.

Table II shows results for log length of stay from versions of Equation (1) with varying sets of controls. All models estimate highly significant and increasingly negative coefficients for approaching time to EOS, with visits seven or more hours prior to EOS being

[16] Because time of arrival at triage is not always observed in the data, I use time at ED floor as my preferred measure of arrival time and condition on pod and providers. The time difference between patient arrival at triage and arrival at ED floor may differ across patient types. For example, patients sent to a 24-hour pod may be sicker than those sent to a partial-day pod; patients sent to a skilled physician may also be sicker than those sent to other physicians. By conditioning on physician and pod identities, I assert that differences in the endogenous time between triage and ED floor that are predictive of patient outcomes are only correlated with physician and pod identities. In practice, however, controlling for physician and pod identities does not matter much for balance.

TABLE I

BALANCE TESTS[a]

	Assumption 1		*Assumption 2*					
	Predicted log LOS	Diagnostic count	Predicted log LOS	Age	ESI	White	Spanish-speaking	Diagnostic count
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Hour prior to EOS								
Last hour	0.003 (0.011)	0.008 (0.070)	−0.008 (0.066)	0.851 (3.297)	−0.004 (0.137)	0.157* (0.092)	0.024 (0.051)	0.140 (0.194)
Second hour	−0.002 (0.006)	−0.028 (0.037)	−0.006 (0.021)	−0.305 (1.088)	−0.001 (0.044)	0.022 (0.030)	−0.006 (0.019)	−0.046 (0.064)
Third hour	0.002 (0.005)	0.001 (0.031)	0.016 (0.015)	0.499 (0.784)	−0.032 (0.031)	0.019 (0.021)	0.004 (0.013)	0.034 (0.046)
Fourth hour	0.004 (0.004)	0.014 (0.026)	−0.006 (0.014)	−0.332 (0.699)	0.000 (0.027)	0.009 (0.019)	0.007 (0.012)	0.014 (0.041)
Fifth hour	0.004 (0.003)	0.011 (0.021)	0.007 (0.013)	0.282 (0.645)	−0.027 (0.025)	−0.022 (0.018)	0.010 (0.011)	0.025 (0.040)
Sixth hour	0.000 (0.003)	−0.010 (0.017)	0.007 (0.012)	−0.075 (0.621)	−0.021 (0.025)	0.000 (0.017)	0.008 (0.010)	0.037 (0.038)
F-test p-value	0.598	0.699	0.852	0.963	0.878	0.339	0.945	0.872
Observations	371,421	371,421	371,421	371,421	371,421	371,421	371,421	371,421
Resid. char. distrib.								
Mean	1.078	1.012	1.079	47.65	2.826	0.495	0.101	1.013
10th percentile	0.985	0.381	0.969	41.69	2.596	0.329	0.017	0.681
90th percentile	1.167	1.623	1.188	53.72	3.058	0.664	0.190	1.356

[a]*Note*: This table assesses balance with respect to two identifying assumptions. Columns 1 and 2 assess Assumption 1, based on Equation (A-1.2), controlling for ex ante patient characteristics time categories, pod, and providers. The remaining columns assess Assumption 2, based on Equation (A-1.3), controlling for time categories, pod, and providers. Further details are given in Appendix A-1.2. Coefficient estimates and standard errors in parentheses are given for each hour prior to end of shift (EOS), where arrival greater than six hours is the reference period. The p-value of the F-test that all coefficients are jointly 0 is also given for each model. The last three rows report summary statistics of the variation in residualized patient characteristics, either across all patients (Columns 1 and 2) or across patient averages at the hourly level (remaining columns). * denotes significance at 10% level. Black race and English-speaking were also assessed for Assumption 2 and similarly yielded null results but are omitted from the table for brevity. Table A-1.1 reports results for other predicted outcomes.

TABLE II

END OF SHIFT EFFECT ON LOG LENGTH OF STAY[a]

	Log length of stay (log hours)							
	OLS (1)	OLS (2)	OLS (3)	OLS (4)	OLS (5)	OLS (6)	Between (7)	Between (8)
Hour prior to EOS								
Last hour	−0.607*** (0.028)	−0.547*** (0.025)	−0.529*** (0.025)	−0.716*** (0.039)	−0.587*** (0.050)	−0.584*** (0.050)	−0.612*** (0.078)	−0.626*** (0.078)
Second hour	−0.316*** (0.008)	−0.282*** (0.008)	−0.330*** (0.008)	−0.461*** (0.012)	−0.287*** (0.026)	−0.287*** (0.026)	−0.325*** (0.033)	−0.344*** (0.035)
Third hour	−0.139*** (0.005)	−0.129*** (0.005)	−0.161*** (0.006)	−0.260*** (0.009)	−0.123*** (0.022)	−0.122*** (0.022)	−0.139*** (0.025)	−0.157*** (0.027)
Fourth hour	−0.112*** (0.005)	−0.092*** (0.004)	−0.111*** (0.005)	−0.173*** (0.008)	−0.091*** (0.018)	−0.089*** (0.018)	−0.098*** (0.021)	−0.107*** (0.022)
Fifth hour	−0.070*** (0.004)	−0.055*** (0.004)	−0.078*** (0.005)	−0.120*** (0.007)	−0.023 (0.015)	−0.022 (0.015)	−0.037** (0.017)	−0.047** (0.019)
Sixth hour	−0.065*** (0.004)	−0.048*** (0.004)	−0.057*** (0.005)	−0.090*** (0.007)	−0.010 (0.012)	−0.010 (0.012)	−0.015 (0.014)	−0.025* (0.015)
Patient characteristics	None	All	All	All	All	Ex ante	All	None
Time and pod	N	N	Y	Y	Y	Y	Y	Y
Physician-resident-nurse identities	N	N	N	Y	Y	Y	Y	Y
Time relative to shift beginning	N	N	N	N	Y	Y	Y	Y
Observations	371,107	371,107	371,107	371,107	371,107	371,107	68,044	68,044
Adjusted R-squared	0.008	0.189	0.211	0.400	0.410	0.404	0.061	0.064
Sample outcome	1.050	1.050	1.050	1.050	1.050	1.050	1.074	1.074

[a]*Note*: This table reports regressions of log length of stay on arrival at each hour prior to end of shift (EOS), where arrival greater than six hours is the reference period. Columns 1 through 6 report results of Equation (1), identified by Assumption 1; Columns 7 and 8 report results of Equation (2), identified by Assumption 2. Patient characteristics include ex ante (to assignment to physician) characteristics of demographics, emergency severity index (ESI), time spent in triage; and ex post Elixhauser indicators for clinical diagnoses. Time dummies include indicators for hour of day, day of week, and month-year interactions. Standard errors are in parentheses and, for "between" models in Columns 7 and 8, are bootstrapped; * denotes significance at 10% level, ** denotes significance at 5% level, and *** denotes significance at 1% level. The adjusted R-squared is given for each overall model in Columns 1 through 6; Columns 7 and 8 report only the adjusted R-squared of the second-stage model in (2) that uses residualized outcomes.

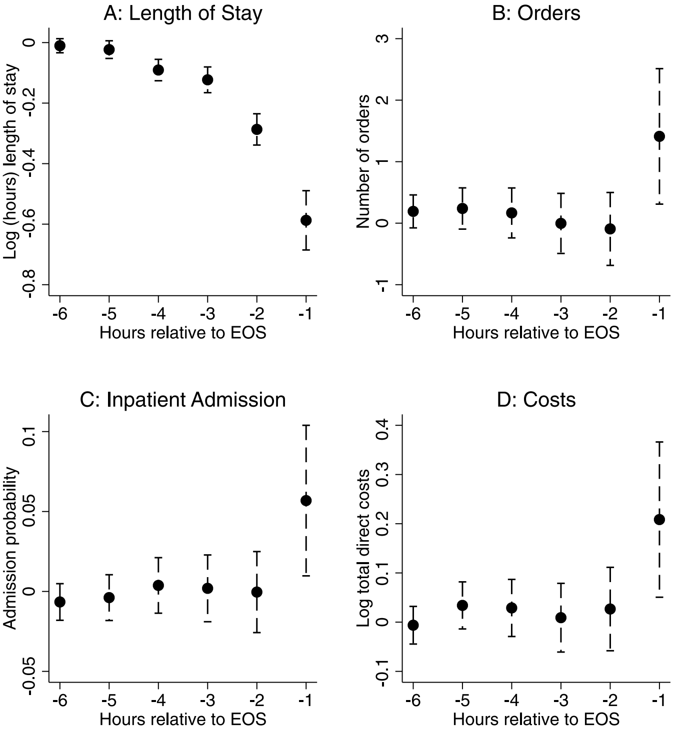

FIGURE 6.—End of shift effects. *Note*: This figure plots average effects for each hour prior to end of shift (EOS) on length of stay (Panel A), orders (Panel B), inpatient admissions (Panel C), and costs (Panel D). Each outcome is estimated separately using Equation (4), and results are the same as in Column 5 in Table II for length of stay and Table III for the other outcomes. The reference category is any time greater than six hours prior to EOS. Bracketed dashed lines represent 95% confidence intervals for each estimate.

the reference category. By the last hour prior to EOS, versions of Equation (1) estimate effects on log length of stay ranging from -0.53 to -0.72. The full model, shown in Column 5 of Table II and plotted in Panel A of Figure 6, estimates an effect on log length of stay of -0.59 in the last hour and serves as the baseline model for this paper.

Results are essentially indistinguishable whether or not all patient characteristics or only ex ante characteristics are included (Columns 5 and 6 in Table II), as suggested by Assumption 1. More generally, results are qualitatively unchanged regardless of whether I control for *any* patient characteristics, time categories, pod dummies, provider identities, or time relative to shift beginning. For example, the difference between Columns 4 and 5 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.[17]

Columns 7 and 8 in Table II report alternative estimates of the EOS effect on length of stay under Assumption 2. Under this assumption, estimates are robust to selection across physicians within hour of arrival and are identified by variation in available shifts (and corresponding hours relative to EOS) *across* hours of arrival. The specification under this

[17]See Appendix A-2 for more direct results on effects relative to shift beginning.

assumption is

$$\overline{Y}_t^* = \sum_{\overline{m}=6}^{1} \alpha_{\overline{m}} P_{\overline{m}}(t) + \varepsilon_t, \tag{2}$$

where $\overline{Y}_t^* \equiv \frac{1}{N_t} \sum_{i \in I(t)} Y_{it}^*$ is the average of residualized length of stay Y_{it}^* for the set $I(t)$ of N_t patients ariving at t, and $P_{\overline{m}}(t)$ is the fraction of these patients being assigned to a shift with $\overline{m}$ hours prior to its end.[18] In this specification, each observation is an arrival hour. Results in Column 7 and 8, which respectively either include or omit *all* patient characteristics, closely match each other and thus support Assumption 2.

In Appendix A-1.4, I adopt an approach by Altonji, Elder, and Taber (2005) to quantify selection on unobservables necessary to explain the observed EOS effects. Using the intuition that estimates change little regardless of controls, I find that normalized selection on unobservables would need to be 475 times greater than normalized selection on observables in order to explain the effect of the last hour before EOS on length of stay.

Table III 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 assigned 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.[19] 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 6.

5. 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 assumption that the EOS by itself has no first-best implications for patient care, conditional on volume of work, time after beginning work, and time after a peer's arrival. Second, this analysis uses shift structure as a concrete example of patient assignment as a policy lever with efficiency tradeoffs: Assigning fewer patients near EOS leaves physicians idle, but assigning more patients worsens the EOS distortion in patient care.

[18]As argued by Chetty, Friedman, and Rockoff (2014), Y_{it}^* is calculated using *within* $\overline{m}$ variation. Section A-1.3 presents an approach and corresponding results that closely follow Chetty, Friedman, and Rockoff (2014) in a way that estimates a single measure of "forecast bias" and additionally accounts for within-shift endogeneity by estimating "jack-knifed" predicted $\hat{\alpha}_{\overline{m}}$ for shift using data that does not include that shift. As suggested elsewhere, results in that section fail to reject the null hypothesis of no forecast bias.

[19]This suggests that formal orders are a net substitute for time. See Appendix A-3 for more direct results supporting this hypothesis.

TABLE III

END OF SHIFT EFFECT ON OTHER OUTCOMES[a]

	Order count (1)	Inpatient admission (2)	Log total cost (3)	30-day mortality (4)	14-day bounce- back (5)	Workload-adjusted log LOS (6)
Hour prior to EOS						
Last hour	1.411** (0.562)	0.057** (0.024)	0.208** (0.080)	−0.003 (0.008)	−0.028 (0.018)	−0.144*** (0.051)
Second hour	−0.093 (0.302)	0.000 (0.013)	0.027 (0.043)	−0.001 (0.004)	−0.011 (0.010)	0.015 (0.027)
Third hour	−0.003 (0.249)	0.002 (0.011)	0.009 (0.036)	−0.005 (0.004)	−0.005 (0.008)	0.090*** (0.022)
Fourth hour	0.167 (0.207)	0.004 (0.009)	0.029 (0.030)	−0.001 (0.003)	−0.002 (0.007)	0.036* (0.018)
Fifth hour	0.239 (0.171)	−0.004 (0.007)	0.034 (0.024)	−0.002 (0.003)	0.001 (0.005)	0.037** (0.015)
Sixth hour	0.192 (0.137)	−0.007 (0.006)	−0.006 (0.019)	0.001 (0.002)	0.001 (0.004)	0.001 (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

[a]*Note*: This table reports coefficient estimates of Equation (1) 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), and workload-adjusted log LOS is the log of this value. Controls are the same as in Column 5 in Table II. Standard errors are in parentheses. * denotes significance at 10% level, ** denotes significance at 5% level, and *** denotes significance at 1% level.

5.1. *Patient Censuses Over Time*

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

$$w(j, t) = \sum_{j(i)=j} \mathbf{1}\big(t \geq t'(i)\big)\mathbf{1}\big(t \leq t'(i) + \tau(i)\big), \tag{3}$$

the count of visits arriving prior to t at $t'(i)$ and staying past t until $t'(i) + \tau(i)$, where $\tau(i)$ is length of stay.

Figure A-7.3 shows unadjusted census averages in 30-minute intervals in different shift types by EOS overlap, $\overline{o}$. Average censuses start at around two patients, representing unstaffed patients from the previous shift, except for shift types with $\overline{o} = 2$, which happen not to transition from another shift (i.e., $\underline{o} = 0$). Patients remain on the census at EOS: Approximately four patients remain on the average census in the last 30 minutes prior to EOS, with the exception of shifts with $\overline{o} = 1$, which have censuses of about six.

5.2. *EOS Effects by Shift Overlap*

I next consider how patient-care EOS effects may differ by shift overlap. With smaller $\overline{o}$, EOS defines earlier times relative to peer arrival after which physicians are *allowed* to go home. Larger patient-care effects with small $\overline{o}$, conditional on time from beginning of shift, are consistent with distortionary care. Further, the interaction reflects an intuitive tradeoff between extensive and intensive margins of distortion: Patient care will be less distorted with larger $\overline{o}$, but this increases slacking off.

I consider three categories $\overline{O}$ of overlap at EOS—terminal shifts ($\overline{o} = 0$), minimally transitioned shifts ($\overline{o} = 1$), and substantially transitioned shifts ($\overline{o} \geq 2$)[20]—and estimate

$$Y_{it} = \big(\alpha^{\overline{O}}_{\overline{m}(i,t)} + \kappa_{\overline{O}}\big)\mathbf{1}\big(\overline{o}(i) \in \overline{O}\big) + \gamma_{\underline{m}(i,t)} + \beta \mathbf{X}_i + \eta \mathbf{T}_t + \zeta_{p(i)} + \nu_{j(i),k(i)} + \varepsilon_{it}, \tag{4}$$

which is similar to Equation (1) but interacts the hourly EOS effects with the categories $\overline{O}$, where $\overline{o}(i)$ is a function assigning visit i to overlap $\overline{o}$ of the shift to which the visit is assigned. In each of the overlap categories, the reference category of $\overline{m}(i, t)$ includes times that are seven hours or more before EOS.

Figure 7 shows interacted EOS effects on length of stay, orders, admission, and total costs. The EOS effect on length of stay is largely similar across shift categories (Panel A). All three shift categories show a substantial decline in length of stay as EOS approaches. However, EOS effects are 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.

5.3. *Effective Time per Patient*

The evidence above suggests 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. To directly assess this concept,

[20] While I observe shifts with $\overline{o} \in \{2, 3\}$, they entail very few observations, as listed in Table A-7.1. Results are essentially unchanged whether I omit these observations or consider them as belonging to the minimally transitioned shift category.

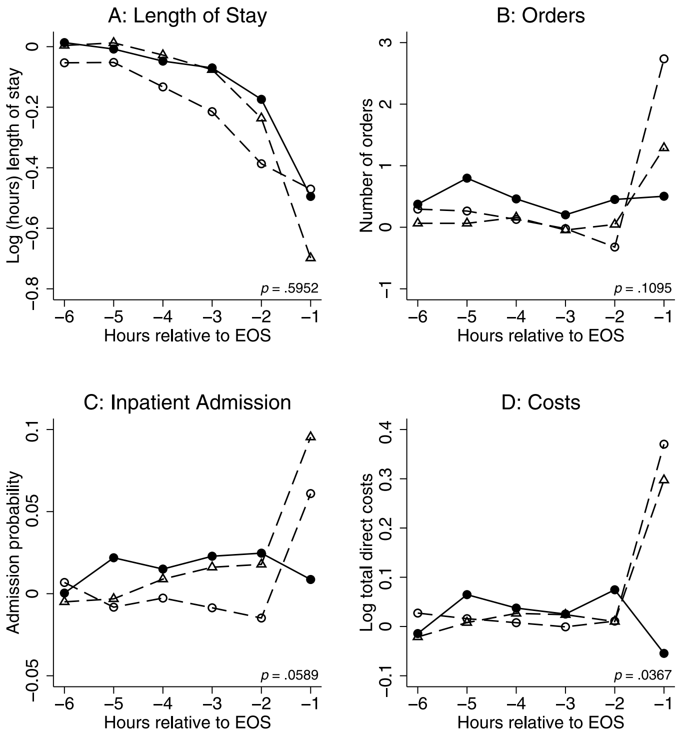

FIGURE 7.—End of shift effects by shift overlap. *Note*: This figure shows end of shift (EOS) effects by EOS overlap times on length of stay (Panel A), orders (Panel B), inpatient admissions (Panel C), and costs (Panel D). Each outcome is estimated separately using Equation (4). Estimates for terminal 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. The noted p-values in each panel are for a test that the coefficient of last hour of shift interacted with $\overline{o} \leq 1$ is different than the coefficient of last hour of shift interacted with $\overline{o} \geq 2$ (these coefficients are slightly different than those shown in the figure in that there are only two categories of shift overlap instead of three).

I create a new outcome measure of *workload-adjusted length of stay*, which normalizes length of stay by the physician's average census during a patient's stay. That is, for visit i arriving at t, I divide length of stay, $\tau(i)$, by the average census $\overline{w}(i)$ under the assigned physician, $j(i)$, over the course of the i's length of stay (from t to $\tau(i)$):

$$\tau(i)/\overline{w}(i) = \tau(i)\left[\frac{1}{\tau(i)}\int_{\tilde{t}=t}^{t+\tau(i)} w\big(j(i), \tilde{t}\big)\, d\tilde{t}\right]^{-1}, \tag{5}$$

where census $w(j, t)$ is defined by Equation (3).

I regress the log of workload-adjusted length of stay using Equation (1).[21] The last column of Table III shows that workload-adjusted length of stay decreases significantly only in the last hour prior to EOS. Thus, adjusting length of stay for workload reconciles previous results in which length of stay progressively decreases as EOS approaches, but orders,

[21]This is different than controlling for current census; results in Table II are unchanged when flexible splines of current census are included in Equation (1). 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.

TABLE IV

EFFECT ON WORKLOAD-ADJUSTED LOG LENGTH OF STAY BY SHIFT OVERLAP[a]

	(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** (0.068)	−0.034 (0.110)	−0.229*** (0.069)	−0.003 (0.143)
Second hour	0.015 (0.038)	0.126* (0.067)	0.014 (0.039)	0.140 (0.085)
Third hour	0.05 (0.031)	0.137** (0.056)	0.037 (0.033)	0.099 (0.069)
Fourth hour	0.007 (0.025)	0.088* (0.051)	−0.002 (0.026)	0.056 (0.058)
Fifth hour	0.013 (0.021)	0.085* (0.044)	0.009 (0.022)	0.052 (0.047)
Sixth hour	−0.017 (0.015)	0.044 (0.038)	-0.022 (0.016)	0.029 (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

[a]*Note*: This table reports coefficient estimates and standard errors in parentheses for EOS effects on workload-adjusted log 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 (1) on subsamples of the data according to $\overline{o}$. All three models are estimated with a full set of controls, as in Column 5 in Table II. Workload-adjusted length of stay (LOS) is calculated by Equation (5), and workload-adjusted log LOS is the log of this value. * denotes significance at 10% level, ** denotes significance at 5% level, and *** denotes significance at 1% level. This table is continued by Table A-7.3.

admissions, and costs increase only in the last hour. In Table IV, I also show that workload-adjusted length of stay is only decreased in the last hour of shift when $\overline{o} \leq 1$. 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.[22] These relationships suggest workload-adjusted length of stay as a relevant measure of time that enters into the physician production function, particularly when considering distortions in orders, admissions, and costs relative to EOS.

5.4. *Time of Shift End*

Finally, I investigate whether EOS effects differ according to the shift end time of day. Specifically, physician leisure is likely to be more valuable at the margin during daytime, when non-work activity may be planned and costlier to defer.[23] I specify two categories $\overline{T}$ of shift end times: shifts with EOS during daytime (from 6:00 a.m. to 8:00 p.m.) and the

[22]If anything, workload-adjusted length of stay slightly increases prior to EOS when $\overline{o} \geq 2$. Such increases do not appear to be associated with changes in other outcomes of orders, admissions, or costs, which could be consistent with increases in length of stay for strategic purposes, or "foot-dragging," as discussed in Chan (2016).

[23]Mas and Pallais (2016) found that workers strongly dislike employers setting schedules on short notice and have a preference for "regular hours," ending at 5:00 p.m. A natural reason for this is that workers value non-work activities that require coordinating with others. Although activities done by oneself, such as sleeping, are valuable, these activities are less costly to defer by one or two hours, if unforeseen circumstances at work occur.

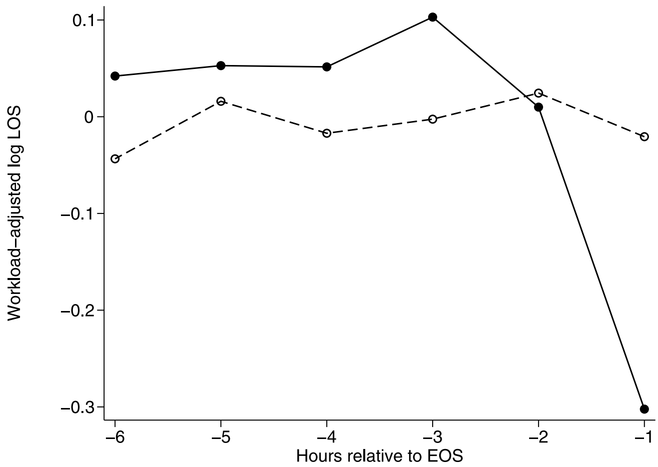

FIGURE 8.—End of shift effect by whether last hour is during daytime. *Note*: This figure shows estimates of effects of hour relative to end of shift (EOS), using Equation (6), conditioning on whether the shift ends during daytime (6 a.m. to 8 p.m.) (solid dots) or not (hollow dots). Roughly half of the shifts end during daytime, with 11,183 shifts ending during daytime and the remainder, 11,344 shifts, ending during any other hour. The p-values of F-tests on the difference between (a) all hourly coefficients corresponding to daytime versus not and between (b) the last-hour of shift coefficients corresponding to daytime versus not are both less than 0.01.

remainder of shifts. I then estimate the following regression:

$$Y_{it} = \left(\alpha^{\overline{T}}_{\overline{m}(i,t)} + \kappa_{\overline{T}}\right)\mathbf{1}\left(\bar{t}(i) \in \overline{T}\right) + \gamma_{\underline{m}(i,t)} + \beta \mathbf{X}_i + \eta \mathbf{T}_t + \zeta_{p(i)} + \nu_{j(i),k(i)} + \varepsilon_{it}, \tag{6}$$

where Y_{it} is log workload-adjusted length of stay, and the reference category of $\overline{m}(i,t)$ includes times that are seven hours or more before EOS.

As shown in Figure 8, the EOS effect on workload-adjusted length of stay is concentrated in shifts ending during daytime but not in other shifts. This is consistent with the benchmark distortion mechanism of leisure-seeking when it is costlier to defer non-work activity. Further, in Figure A-7.2, I show that physicians are less likely to continue to write orders past EOS in shifts ending during daytime, despite higher remaining patient loads at EOS in these shifts.

6. STYLIZED MODEL OF OPTIMAL ASSIGNMENT

I introduce a simple model to consider how physician decisions may be distorted under work schedules. The key distortionary elements of the model are the following: (1) Workers have private information about their tasks; (2) workers care more about their own income and leisure relative to the social consequences of their actions; and (3) ex post worker-task specificity prevents workers from simply passing off tasks at EOS (Briscoe (2007), Goldin (2014)). While the assignment of tasks (patients) is observable, subsequent performance on the tasks is not contractible. This implies a second-best assignment policy that takes this into account.

6.1. *Model Setup*

Consider a physician in a shiftwork arrangement: She has a contract to work on a shift 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 are captured by $\mathcal{E}_t$, which includes, among other things, the start time of the physician's shift, her workload, w_t, and the start time and workload of a potential peer who may take the patient instead. The patient's underlying health state, $\theta \in \{0, 1\}$ for whether the patient is healthy ($\theta = 0$) or sick ($\theta = 1$), is unknown at this point, but $\Pr(\theta = 1) = p$ is publicly known. The timing is as follows:

1. The physician may be assigned the patient ($a = 1$) or not ($a = 0$).
2. If she is assigned the patient, she observes private information $\mathscr{I}$ so that $\Pr(\theta = 1|\mathscr{I}) = p'$, and $|p' - \theta| < |p - \theta|$.[24] She decides on patient care inputs: time τ and formal tests and treatments z.
3. The physician observes $\hat{\theta} = \theta$ with probability $q(\tau, z) \in (0, 1)$ and no information ($\hat{\theta} = \emptyset$) with probability $1 - q(\tau, z)$. She decides $d \in \{0, 1\}$, 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:

$$U = \begin{cases} y + \lambda O(\theta; \mathcal{E}_t), & a = 0, \\ y - \tilde{c}_\tau(\tau) + \lambda\big(v(\theta, d) - c(\tau, z)\big), & a = 1. \end{cases} \tag{7}$$

Utility is stated in dollar terms, where physician income y does not depend on her actions.[25] $O(\theta; \mathcal{E}_t)$ 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 discharge decision d for patient with health θ. $c(\tau, z)$ 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 (7), 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).

6.2. *Patient Care*

I first examine EOS effects on patient care and the discharge decision, conditional on assignment ($a = 1$). 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: $v(1, 1) - v(1, 0) > v(0, 0) - v(0, 1)$.[26] Because of this last fact, if θ remains unobserved, the physician will admit if

[24] I rule out private information before patient acceptance in this model. This is generally consistent with the institutional setting, and I examine this empirically as Assumption 1 below.

[25] In scheduled work y mostly depends on 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.

[26] I assume that the physician values discharge actions relative to the patient's health state the same as the social planner does. An additional wedge of *defensive medicine* would further increase the *physician's* $v(1, 1) - v(1, 0)$ relative to $v(0, 0) - v(0, 1)$ but not the social planner's valuations. An EOS distortion in patient care, in concert with this distortion, would likely be larger, given concavity in $q(\tau, z)$ described formally below.

and only if $p' > p^*$, where $p^* < \frac{1}{2}$.[27] In other words, the realized discharge decision d is a function of $\hat{\theta}$:

$$d(\hat{\theta}) = \begin{cases} \theta, & \hat{\theta} = \theta, \\ \mathbf{1}(p' > p^*), & \hat{\theta} = \emptyset. \end{cases}$$

The probability q of observing θ is in turn a function of patient care inputs (τ, z).[28] $q(\tau, z; w_t)$ is increasing and concave with respect to both τ and z and also depends on workload of the physician w_t. 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 $\partial^2 q/(\partial z\,\partial w_t) = 0$. 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.[29] By the normalizing assumption, I focus attention on substitutability or complementarity between time and formal utilization. Costs in $c(\tau, z)$ and $\tilde{c}_\tau(\tau)$ are positive, continuous, increasing, and convex in their arguments. Define $\tilde{c}_\tau = 0$ for $\tau + t - \overline{t} \leq 0$.

PROPOSITION 1: *Denote inputs in Section 6.1 that maximize expected utility in Equation (7), conditional on patient assignment* $(a = 1)$, *as* $(\tau^*(t), z^*(t))$. *Denote corresponding inputs that maximize welfare as* $(\tau^{\mathrm{FB}}(t), z^{\mathrm{FB}}(t))$. *Assume that* $F_{p'}(p^*) < \frac{1}{2}$.

(a) *As* $t \to \overline{t}$, $\tau^*(t)$ *weakly decreases*, $z^*(t)$ *may weakly increase (if* τ *and* z *are net substitutes) or decrease (if* τ *and* z *are net complements), and* $E[d|\tau^*(t), z^*(t)]$ *weakly increases*.

(b) *For all* t, $\tau^*(t) \leq \tau^{\mathrm{FB}}(t)$, *and* $E[d|\tau^*(t), z^*(t)] \geq E[d|\tau^{\mathrm{FB}}(t), z^{\mathrm{FB}}(t)]$.

(c) *If* τ *and* z *are net substitutes, then* $z^*(t) \geq z^{\mathrm{FB}}(t)$ *for all* t, *and* $z^*(t) - z^{\mathrm{FB}}(t)$ *weakly increases in* w_t, *holding* t *constant. The reverse is true if* τ *and* z *are net complements.*

PROOF: See Appendix A-4.1. *Q.E.D.*

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 $a^* = 1$. Note that distortions in discharges are due solely through distortions in $(\tau^*(t), z^*(t))$, which lower q: conditional on (τ, z), physicians make the right decision on d.

6.3. *Patient Assignment*

I next consider the case in which the physician is allowed to maximize her utility by choosing whether to accept the new patient, $a \in \{0, 1\}$. The physician will accept the pa-

[27]This can be straightforwardly shown by noting that $E[v(\theta, 0)|\Pr(\theta = 1) = p^*] = E[v(\theta, 1)|\Pr(\theta = 1) = p^*]$.

[28]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.

[29]In its most straightforward form, effective time per patient can be viewed as the concept behind "workload-adjusted length of stay," in Section 5.3, which is length of stay divided by the average number of patients on census during a given patient's length of stay. This concept reflects that the cost of a patient's length of stay as an input does not depend on how many other patients a physician concurrently cares for, but the effectiveness of this time as an input into q is reduced by workload.

tient if and only if the expected utility of the outside option, $E[O(\theta; \mathcal{E}_t)]$, is below some threshold $\underline{O}^*$. We can define similar thresholds that maximize welfare in the first-best case in which the social planner can implement (τ^{FB}, z^{FB}) and in the second-best case in which the social planner has no knowledge of appropriate patient-care inputs, denoted as $\underline{O}^{FB}$ and $\underline{O}^{SB}$, respectively.

PROPOSITION 2: *Consider a^* as the patient assignment in Section 6.1 that maximizes expected utility in Equation (7), a^{FB} as the assignment that maximizes expected welfare when optimal (τ^{FB}, z^{FB}) is publicly known and contractible, and a^{SB} as the assignment that maximizes expected welfare when (τ^{FB}, z^{FB}) is either publicly unknown or non-contractible. Assignment will follow threshold rules in which assignment occurs if and only if $E[O(\theta; \mathcal{E}_t)]$ 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.*

PROOF: See Appendix A-4.2. *Q.E.D.*

There are obvious reasons for even first-best assignment to decrease near EOS: As $t \to \bar{t}$, the outside option $O(\theta; \mathcal{E}_t)$ increases because a peer is more likely to be arriving soon or already present, and welfare-relevant costs of producing q change as the physician approaches EOS. Beyond this, however, assignment decided by the physician, a^*, will be lower than the first-best assignment because she overvalues her leisure relative to other welfare-relevant concerns (i.e., $\underline{O}^* < \underline{O}^{FB}$). In the second-best policy, the social planner maximizes welfare with patient assignment, but under the constraint that the physician still has control of inputs and will continue to choose (τ^*, z^*). The second-best threshold $\underline{O}^{SB}$ will be higher than $\underline{O}^*$, because the social planner does not overvalue the physician's leisure. But $\underline{O}^{SB}$ will be lower than $\underline{O}^{FB}$, because overvalued leisure still distorts patient care, and the social planner cannot implement appropriate patient care. The greater the patient-care distortions, the closer $\underline{O}^{SB}$ will be to $\underline{O}^*$ than to $\underline{O}^{FB}$.

7. COUNTERFACTUAL ASSIGNMENT POLICIES

Because the assignment of patients is observable, it is natural to conceive ex ante policies of patient assignment, while downstream patient-care decisions are non-contractible because patient types are not in general publicly observed. In Section 6, I show in a highly stylized conceptual framework that the second-best policy involves some slacking off. In this section, I take this concept to the data, simulating outcomes under counterfactual assignment policies along the margin of slacking off. I consider patient assignment as a sufficient statistic for a wider range of managerial policies (e.g., rules, shift overlap, financial incentives) that may influence how patients are assigned but do not separately specify or incentivize how patients are to be treated.

7.1. *Simulation Approach*

At a high level, there are three components of simulating outcomes under counterfactual assignment policies. First, I specify and estimate a model of physician discharge choices over multiple patients as a dynamic discrete choice problem. Second, I consider counterfactual assignment policies, and using the dynamic programming framework, determine physician discharge probabilities under these counterfactuals. Third, I simulate assignments and discharges under each counterfactual, and use these simulated observations to impute welfare-relevant costs of physician time, patient time, and hospital resources. I provide further details in Appendix A-5.

Consider an ED physician j at time t in state S, which includes shift characteristics that the physician is working in, time categories (e.g., time of day, day of week), time relative to EOS, and the set of patients (and their characteristics) already under her care or the care of another physician present. I model the doctor's decision to discharge patients, conditional on S, as a dynamic discrete choice problem. In short time intervals, I assume that doctor may discharge at most one patient under her care, $i \in \mathcal{I}$ (including $i = \emptyset$, or no discharge of any patient), and this decision may be represented as a Bellman equation:

$$V(S) = E\left[\max_{i \in \mathcal{I}} \left\{ u(i, S) + \delta \int_{S'} V(S')\, dF(S'|i, S) \right\}\right], \tag{8}$$

where $V(S)$ is the value function of the dynamic programming problem, $u(i, S)$ is the stochastic utility flow of choosing i in state S, δ is the discount factor, and $F(S'|i, S)$ is the Markov transition distribution function. The physician chooses i that maximizes the stochastic utility flow and the expected value function in the next period. I specify the flow utility as

$$u(i, S) = \overline{u}(i, S) + \varepsilon_i,$$

where $\overline{u}(i, S) = \mathbf{b}(i, S)\theta_u$ is a linear combination of basis splines $\mathbf{b}(i, S)$, and ε_i is an i.i.d. Type I extreme value shock. $\mathbf{b}(i, S)$ are splines of patient i's predicted length of stay and of the deviation in current length of stay from this prediction. Importantly, I exclude time to EOS from entering into $\overline{u}(i, S)$.

I first estimate the transition function, $F(S'|i, S)$, focusing on uncertainty in the assignment of new patients. I estimate the number of patients assigned to a physician in state S using an ordered logit model and each patient's predicted length of stay using OLS. To estimate θ_u in the utility flow function, I use tools from the dynamic discrete choice literature. Specifically, the log-likelihood of observing actual patient discharges (i.e., indicators $d(i, t)$ for whether patient i was discharged by physician j at time t) is

$$\log \mathcal{L} = \sum_{j,t} \sum_{i \in \mathcal{I}(j,t)} d(i, t) \log \widehat{\Pr}\big(d(i, t)|i, S_{j,t}\big),$$

where $\widehat{\Pr}(d(i, t)|i, S_{j,t})$ is the implied conditional choice probability under parameters θ_u. Dropping subscripts for simplicity, the probability of discharge choice i conditional on S is

$$\Pr(i|S) = \frac{\exp\big(\overline{v}(i, S)\big)}{\sum_{i' \in I} \exp\big(\overline{v}(i', S)\big)}, \tag{9}$$

where $\overline{v}(i, S) = \overline{u}(i, S) + \delta \int_{S'} V(S')\, dF(S'|i, S)$. If S were in a low-dimensional space, I could solve for θ_u using a nested fixed-point algorithm. However, because S is high-dimensional, I use a constrained optimization approach proposed by Barwick and Pathak (2015), building on the Mathematical Program with Equilibrium Constraints (MPEC) concept proposed by Su and Judd (2012). Rather than solving for $V(S)$ using backward induction, this approach approximates $V(S)$ using splines and incorporates a Bellman equation identity as constraints for the maximum likelihood problem.

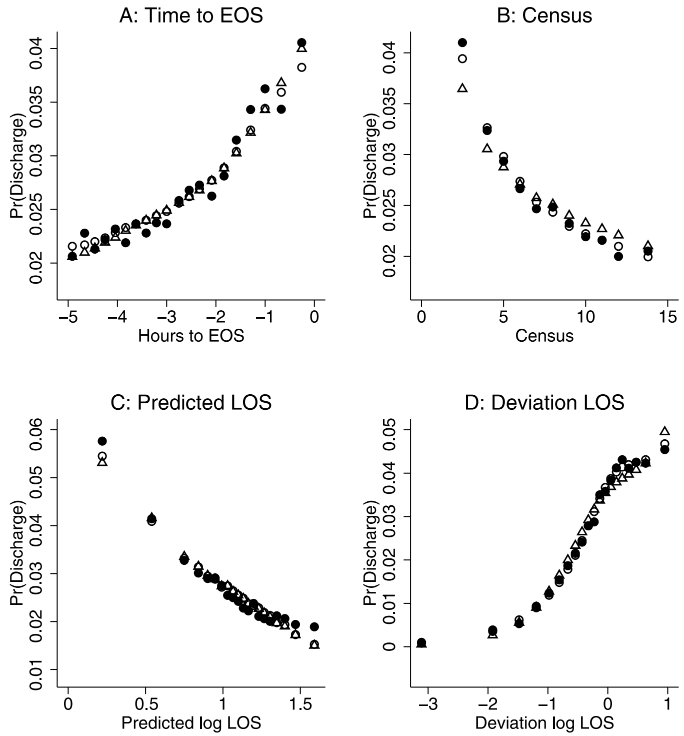

FIGURE 9.—Model fit by discharge probabilities. *Note*: This figure shows the fit of the discharge policies with actual discharge decisions in a sample of 792,687 patient-time observations. Each panel evaluates the fit between the actual data (solid dots), a flexible multinomial logit "static" model of discharges (hollow dots), and a dynamic model of discharges with restrictions on what can enter into the utility flow (triangles). The dynamic model corresponds to the simplest model in Table A-5.1, Model 1, although the fit and the estimated parameters do not qualitatively differ across specifications. The y-axis in all panels is the probability or average of whether a patient i still undischarged as of time $t-1$ could be discharged at t by doctor j on shift with certain state characteristics. The x-axis is some characteristic of (i, j, t): hours that (j, t) is away from EOS in Panel A; the number of patients on census, or $\|I(j, t)\|$, in Panel B; predicted length of stay according to i's characteristics and time categories of t in Panel C; the deviation in current log length of stay from the predicted log length of stay in Panel D. Each marker potentially represents 1/40th of the data if the x-axis is sufficiently continuous. The dynamic programming approach is described in greater detail in Appendix A-5.1.

Estimates of the underlying utility flow function, $\overline{u}(i, S)$, are robust across functional form specifications of both $\overline{u}(i, S)$ and $V(S)$; parameter results are presented in Table A-5.1. Discharge probabilities implied by the model fit the underlying data well, as shown in Figure 9. Appendix A-5.1 provides further details on identification, estimation, robustness, and goodness of fit of the dynamic programming problem.

With estimates of θ_u in hand, the exercise is to consider counterfactual assignment policies that change $F(S'|i, S)$. I consider counterfactual assignment policies along the margin of slacking off, adjusting the way time to EOS is considered, as parameterized by a scalar Δ measured in hours, further described in Appendix A-5.2. If $\Delta < 0$, then fewer patients will be assigned; if $\Delta > 0$, then more patients will be assigned. Under a counterfactual $F(S'|i, S)$, I recompute the Bellman equation, which gives conditional choice probabilities by Equation (9). Physicians may dynamically respond to counterfactual assignment policies, for example by further delaying the discharge of patients to forestall new patient assignments if new assignments are more likely near EOS (Chan (2016)).

Given counterfactual assignment policies and conditional choice probabilities for each Δ, I simulate patient assignments and discharges. I then use these counterfactual sequences of assignments and discharges to impute welfare-relevant costs of physician time, patient time, and hospital resources, for each counterfactual Δ and simulation $r = 1, \dots, 20$:[30]

$$\text{Costs}_{\Delta,r} = \text{PhysicianTime}_{\Delta,r} + \text{PatientTime}_{\Delta,r} + \text{HospitalResources}_{\Delta,r}. \tag{10}$$

$\text{PhysicianTime}_{\Delta,r}$ captures additional wages that the ED must pay in order to meet patient flow and is calculated as the number of physician hours needed to overall patient flow. Two considerations may increase $\text{PhysicianTime}_{\Delta,r}$. First, if fewer patients are assigned before EOS, a peer must arrive earlier because backlog occurs earlier. On the other hand, if more patients are assigned prior to EOS, the index physician must stay later past EOS, and this foregone leisure is also valuable. I value physician time with a base-case wage of \$120/hour, which is close to actual wages in this ED and national averages of hourly pay. To reflect that, all else equal, patients prefer shorter stays, I calculate $\text{PatientTimeCosts}_{\Delta}^{r}$ as the total time spent by patients in the ED, valued at \$20/hour. I calculate $\text{HospitalResources}_{\Delta}^{r}$ by using previously estimated causal effects on length of stay and hospital resources, in Sections 4 and 5. Under an additional assumption that hospital resources are affected via the EOS mechanism of reducing lengths of stay, these two effects imply a marginal rate of technical substitution between time and hospital resources. This is nonparametrically identified, although based on results in Section 5.3, I choose to use workload-adjusted length of stay as the relevant measure impacting hospital costs.[31] Further details are given in Appendix A-5.3.

7.2. *Results*

Increasing assignment near EOS results in large increases in hospital-resource costs, which dominate any physician-time savings from having the next physician arrive later. For example, an assignment policy that results in physicians staying an extra hour past EOS by assigning them about four additional patients also results in an increase of \$3700 in hospital spending per shift, for a net increase of \$3300 in total costs, compared to a policy that assigns those four patients to a physician who is not ending shift. Figure 10 shows average changes in total costs per shift, defined by Equation (10), under counterfactual policies, where policies are shown in terms of changes in the number of patients assigned.[32] The observed pattern of assignment—implemented by shift overlap, triage

[30]This makes the conservative assumption that patient health is unaffected despite EOS distortions in time, formal utilization, and admissions. In sample, recall that I find no effect on mortality or bounce-backs (Table III). Out of sample, with greater assignments near EOS, the EOS distortion may also worsen health outcomes, which implies even greater importance to reduce assignments near EOS. However, this should not matter for Δ close to 0 and therefore for statements about optimal assignment if the optimal assignment policy occurs close to the observed assignment regime.

[31]The elasticity estimate is motivated by the fact that both observed total costs (Figure 6) and observed workload-adjusted length of stay (Figure A-7.4) increase only in the last hour prior to EOS, in Section 5.2. 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-7.3. 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 detail is given in Appendix A-5.4.

[32]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., $|\overline{m}(\Delta) - \overline{m}| \leq \Delta$). This can be appreciated in plots of curtailed assignment policies in Figure A-5.1, in which $\Delta = -2$ is still very similar to $\Delta = 0$.

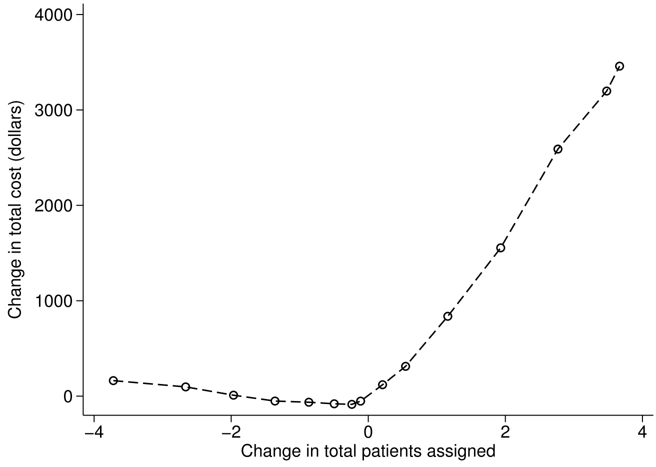

FIGURE 10.—Simulated change in total cost per shift under counterfactual assignment. *Note*: This figure plots simulated changes in total cost for each counterfactual assignment regime, as described in Section 7 and Appendix A-5. Assignment policies may either increase or decrease assignment, as illustrated in Figure A-5.1. Counterfactual costs depend on a dynamic programming model of physician discharge decisions. These results are generated from parameters of the dynamic programming model corresponding to Model 1 in Table A-5.1, although results do not qualitatively differ across model specifications. The x-axis is the change in total patients assigned per shift as a result of an assignment regime ("0" corresponds to the actual assignment regime). Daily costs include physician-time, patient-time, and hospital-resource costs. Each marker is an average over 20 simulations in an assignment policy $\mathcal{A}_\Delta$, and each simulation contains 280,000 to 460,000 observations, depending on whether $\mathcal{A}_\Delta$ decreases or increases the number of patients assigned relative to the actual assignments.

nurse assignment, and physician responses—approximately minimizes total costs relative to counterfactual policies, implying that it is close to second-best optimal. Because hospital resources are potentially costlier, increasing assignments increases costs more dramatically than does reducing assignments.

The intuition behind this robust finding is that, in the actual assignment regime $\mathcal{A}_0$, very few patients are assigned in the last hour of shift, and this is precisely when costs start increasing. Reducing assignment near EOS would not substantially change in absolute terms the number of patients with EOS-induced distortions in hospital resources, but of course would increase physician time to cover shift changes. On the other hand, increasing assignment near EOS is much costlier for two reasons. First, patient care costs are much higher than physician wages. Second, increasing assignment not only exposes more patients to EOS cost distortions but also worsens the distortion per patient if the relevant time input measure is time per patient (e.g., workload-adjusted length of stay). Thus, if a planner were uncertain about when the optimal assignment policy should begin reducing assignments, it would be cheaper and less risky to pay for idle physician time as opposed to risking patient-care distortions.

Another way to use the 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 hospital-resource costs incurred per leisure hour gained, shown in Figure 11 (details in Appendix A-5.4.1). At times prior to normal completion (but after EOS), the value of incremental leisure appears low, below the market wage of $120 per hour. However, the value of leisure quickly rises above market wage as work completion time increases. At about one hour past nor-

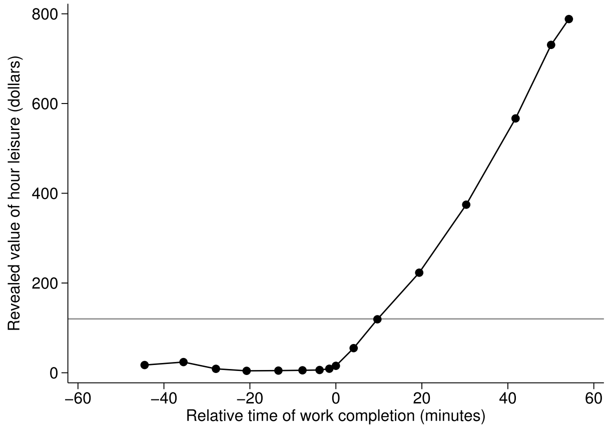

FIGURE 11.—Value of leisure in dollars of patient care. *Note*: This figure plots the imputed value of leisure, revealed by increases in resource-utilization costs that shorten the time for completion of work under simulated counterfactual assignment policies. For each counterfactual assignment policy, data are simulated using the baseline counterfactual discharge policy $\mathcal{D}_\Delta$ and then again for a counterfactual discharge policy $\underline{\mathcal{D}}_\Delta$ 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. These results are generated from parameters of the dynamic programming model corresponding to Model 1 in Table A-5.1, although results do not qualitatively differ across model specifications. The x-axis is the relative time of work completion, compared to assumed work completion under the observed assignment and discharge policies, $\mathcal{A}_0$ and $\mathcal{D}_0$, respectively. 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-5.4.1.

mal completion, physicians are willing to expend $790 in hospital resources, six times above the market wage, in order to avoid an additional hour at work.[33]

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)).[34] Slacking off has been described as tapering work, particularly in the context of shirking near the end of scheduled work.[35] Both of these concepts are closely related

[33]Under a strict interpretation of Equation (7) and no (assumed) worsening of patient health, this implies that λ is decreasing after EOS and that $\lambda = \$120/\$790 = 0.15$ at this point. In another dimension of heterogeneity, in Section 5.4 and Figure 8, I show reduced-form evidence that the EOS effect on workload-adjusted length of stay is entirely in shifts ending during daytime hours between 6:00 a.m. and 8:00 p.m., which is consistent with leisure being more highly valued during these times, when others are awake and non-work activities are costlier to defer.

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

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

to the phenomenon described in this paper. Despite the negative connotation of these terms, I argue that informational frictions (i.e., when work can be assigned but not much else is observable or contractible) imply some slacking off would be optimal in a second-best sense. This may explain why the practice—either tapering assignments or equivalently scheduling time at work to be longer than necessary—is not only prevalent but also tolerated. In the setting of this ED, I find that the observed assignment policy is in fact approximately (second-best) optimal.

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. For example, social norms may pressure workers not to 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.[36] Workers may take schedules as a contractual "reference point" (Hart and Moore (2008)), and there may even be accepted routines (e.g., sign-out rounds) that reinforce this sense. All of these mechanisms would have equivalent implications on how availability should be scheduled and worked assigned, but some mechanisms (e.g., social or rule-based norms) could in principle be directly tackled by an organization.[37]

Workers in practice may underestimate the time it takes to complete work, paying too much attention to when EOS is officially set when planning after-work activities. Even rational workers would view uncertainty with respect to the time they leave work as costly, if they have plans that cannot be easily changed in the short term, such as picking up children from school or having a dinner reservation. This implies that the canonical model of a constant value of leisure is misspecified, and that short-term reductions in "leisure" are much costlier than prearranged work, as has been noted in the press with practices such as "flexible" or "just-in-time" scheduling that benefit employers but could be costly for employees (e.g., Greenhouse (2014)). Accounting for this would imply even greater welfare benefits from slacking off to ensure that workers leave close to a prespecified time.

Eliminating Transfer Costs. Without transfer costs between physicians, there would be no EOS distortion. The results from simulating counterfactual assignment policies, for example, that assigning merely four additional patients would increase net costs for a shift by \$3300, suggest that potential transfer costs are quite large, at least in terms of hospital resources. Transfer costs could reflect technological difficulties in passing information to subsequent physicians or could simply reflect behavioral norms that discourage transfers.[38] A hypothetical scenario in which transfer costs are eliminated could fur-

[36]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. In Appendix A-6 (Figures A-6.2 and A-6.1), I show that EOS distortions are more likely under certain physician types and physician-peer relationships, for example, slower physicians and physicians handing off to a senior peer, a faster peer, or a peer with whom they are unfamiliar.

[37]In Figure A-7.1, I show evidence that physicians continue to write orders past EOS, consistent with the qualitative report from this ED that most physicians stay two to three hours past EOS (including activities such as dictating notes, conferring with other staff, and notifying patients, in addition to order-writing). In addition, Section 5.4 discusses evidence that EOS effects and reductions in attending-physician order-writing are both concentrated in shifts ending during daytime, when non-work activity is costlier to defer. This is consistent with the benchmark distortion mechanism of leisure-seeking.

[38]Transfer costs have been discussed in the economics literature as "worker-task specificity" (e.g., Goldin (2014)) and in the medical literature manifesting as difficulties in passing information between providers tran-

ther improve welfare by removing the need to slack off. However, it appears that welfare improvements in this hypothetical scenario would be relatively limited for two reasons: (i) very few patients are affected by increased hospital costs under the actual assignment policy, so reductions in hospital costs would be minimal; and (ii) the cost of an hour or two of physician time is (relatively) cheap, so reducing slacking off would not save much. Further, there are likely general behavioral foundations for transfer costs that could prove difficult to eliminate.[39]

Schedules, Prices, and Spending Budgets. 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 superior when benefits are more concave than costs are convex, which characterizes production within most organizations at least in the short term (Weitzman (1974)).

This reasoning likely underlies the very existence of schedules. For example, in the ED, given uncertain work flow, it could be very costly to simply post wages *in advance* and hope that the appropriate number of physicians show up for the appropriate number of hours, as opposed to specifying the number of hours each physician should be there. With asymmetric information and moral hazard, price or budget mechanisms could be 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 $c(\tau, z)$ are appropriately weighted), then a global budget could distort care uniformly toward underprovision. Instead, simple contracts that allow physicians to treat the neediest patients first are likely to be the least distortionary.

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 costlier, 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.

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

sitioning care (e.g., Apker, Mallak, and Gibson (2007)). As an example of information-driven transfer costs, a physician may find it difficult to instruct another physician to "discharge the patient home if his signs and symptoms are $x\%$ better," if this information is multidimensional and understood only by personal examination and questioning. Communicating everything the physician knows about a patient is generally difficult. As an example of behavioral-driven transfer costs, physicians may not want to burden a colleague even if it is efficient to do so.

[39]Polanyi (1966) has described tacit knowledge that is difficult to communicate, which could apply to complex patient care, at least in the sense that successful communication is costly both for the sender and the receiver (Dewatripont and Tirole (2005)). Although patient care shared with nurses and residents who overlap with attending physicians changing shifts could reduce transfer costs, the extent of this is also limited by knowledge and communication issues with attending physicians responsible for the discharge decision.

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” appear approximately second-best optimal, as I simulate large welfare losses in assignment policies without tapering near EOS. 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 (1997): “Formal and Real Authority in Organizations,” *The Journal of Political Economy*, 105, 1–29. [999]

AKERLOF, G. A., AND R. E. KRANTON (2005): “Identity and the Economics of Organizations,” *The Journal of Economic Perspectives*, 19, 9–32. [999]

ALTONJI, J. G., T. E. ELDER, AND C. R. TABER (2005): “Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools,” *Journal of Political Economy*, 113, 151–184. [1012]

APKER, J., L. A. MALLAK, AND S. C. GIBSON (2007): “Communicating in the Gray Zone: Perceptions About Emergency Physician-Hospitalist Handoffs and Patient Safety,” *Academic Emergency Medicine*, 14, 884–894. [1002,1027]

BANDIERA, O., I. BARANKAY, AND I. RASUL (2005): “Social Preferences and the Response to Incentives: Evidence From Personnel Data,” *The Quarterly Journal of Economics*, 120, 917–962. [999]

——— (2009): “Social Connections and Incentives in the Workplace: Evidence From Personnel Data,” *Econometrica*, 77, 1047–1094. [999]

BARWICK, P. J., AND P. A. PATHAK (2015): “The Costs of Free Entry: An Empirical Study of Real Estate Agents in Greater Boston,” *The RAND Journal of Economics*, 46, 103–145. [1021]

BEERS, T. (2000): “Flexible Schedules and Shift Work: Replacing the 9-to-5 Workday?” Tech. Rep., Bureau of Labor Statistics. [998]

BENABOU, R., AND J. TIROLE (2003): “Intrinsic and Extrinsic Motivation,” *The Review of Economic Studies*, 70, 489–520. [999]

BESLEY, T., AND M. GHATAK (2005): “Competition and Incentives With Motivated Agents,” *American Economic Review*, 95, 616–636. [999]

BLOOM, N., J. LIANG, J. ROBERTS, AND Z. J. YING (2015): “Does Working From Home Work? Evidence From a Chinese Experiment,” *The Quarterly Journal of Economics*, 130, 165–218. [999]

BRISCOE, F. (2006): “Temporal Flexibility and Careers: The Role of Large-Scale Organizations for Physicians,” *Industrial and Labor Relations Review*, 60, 88. [1000]

——— (2007): “From Iron Cage to Iron Shield? How Bureaucracy Enables Temporal Flexibility for Professional Service Workers,” *Organization Science*, 18, 297–314. [1017]

CASALINO, L. P., K. J. DEVERS, T. K. LAKE, M. REED, AND J. J. STODDARD (2003): “Benefits of and Barriers to Large Medical Group Practice in the United States,” *Archives of Internal Medicine*, 163, 1958–1964. [1000]

CHAN, D. (2016): “Teamwork and Moral Hazard: Evidence From the Emergency Department,” *Journal of Political Economy*, 124, 734–770. [1016,1022]

CHAN, D. C. (2018): “Supplement to ‘the Efficiency of Slacking off: Evidence From the Emergency Department’,” *Econometrica Supplemental Material*, 86, https://doi.org/10.3982/ECTA13565. [1006]

CHETTY, R., J. N. FRIEDMAN, AND J. E. ROCKOFF (2014): “Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates,” *American Economic Review*, 104, 2593–2632. [1012]

COVIELLO, D., A. ICHINO, AND N. PERSICO (2014): “Time Allocation and Task Juggling,” *American Economic Review*, 104, 609–623. [998]

DEWATRIPONT, M., AND J. TIROLE (2005): “Modes of Communication,” *Journal of Political Economy*, 113, 1217–1238. [1027]

DEWATRIPONT, M., I. JEWITT, AND J. TIROLE (1999): “The Economics of Career Concerns, Part II: Application to Missions and Accountability of Government Agencies,” *The Review of Economic Studies*, 66, 199–217. [999]

ELIXHAUSER, A., C. STEINER, D. R. HARRIS, AND R. M. COFFEY (1998): “Comorbidity Measures for Use With Administrative Data,” *Medical Care*, 36, 8–27. [1004]

ELLIS, R. P., AND T. G. MCGUIRE (1986): "Provider Behavior Under Prospective Reimbursement: Cost Sharing and Supply," *Journal of Health Economics*, 5, 129–151. [999]
FORSTER, A. J., I. STIELL, G. WELLS, A. J. LEE, AND C. VAN WALRAVEN (2003): "The Effect of Hospital Occupancy on Emergency Department Length of Stay and Patient Disposition," *Academic Emergency Medicine*, 10, 127–133. [1004]
GARICANO, L. (2000): "Hierarchies and the Organization of Knowledge in Production," *Journal of Political Economy*, 108, 874–904. [997]
GOLDIN, C. (2014): "A Grand Gender Convergence: Its Last Chapter," *American Economic Review*, 104, 1091–1119. [1000,1017,1026]
GREEN, L. (1984): "A Multiple Dispatch Queueing Model of Police Patrol Operations," *Management Science*, 30, 653–664. [997]
GREEN, L. V. (2004): "Capacity Planning and Management in Hospitals," in *Operations Research and Health Care*. International Series in Operations Research & Management Science, Vol. 70, ed. by M. L. Brandeau, F. Sainfort, and W. P. Pierskalla. Springer US, 15–41. [997]
GREENHOUSE, S. (2014): "Part-Time Schedules, Full-Time Headaches," *The New York Times*. [1026]
HART, O., AND B. HOLMSTROM (1987): "The Theory of Contracts," in *Advances in Economic Theory*, ed. by T. F. Bewley. Cambridge, UK: Cambridge University Press, 71–156. [999]
HART, O., AND J. MOORE (2008): "Contracts as Reference Points," *The Quarterly Journal of Economics*, 123, 1–48. [1026]
HE, B., F. DEXTER, A. MACARIO, AND S. ZENIOS (2012): "The Timing of Staffing Decisions in Hospital Operating Rooms: Incorporating Workload Heterogeneity Into the Newsvendor Problem," *Manufacturing and Service Operations Management*, 14, 99–114. [997]
KATZ, L., AND A. KRUEGER (2016): "The Rise and Nature of Alternative Work Arrangements in the United States, 1995–2015," Tech. Rep. w22667, National Bureau of Economic Research, Cambridge, MA. DOI:10.3386/w22667. [998]
KC, D. S., AND C. TERWIESCH (2009): "Impact of Workload on Service Time and Patient Safety: An Econometric Analysis of Hospital Operations," *Management Science*, 55, 1486–1498. [997]
KOLSTAD, J. T. (2013): "Information and Quality When Motivation Is Intrinsic: Evidence From Surgeon Report Cards," *American Economic Review*, 103, 2875–2910. [999]
LARKIN, I. (2014): "The Cost of High-Powered Incentives: Employee Gaming in Enterprise Software Sales," *Journal of Labor Economics*, 32, 199–227. [999]
LERMAN, B., AND M. S. KOBERNICK (1987): "Return Visits to the Emergency Department," *The Journal of Emergency Medicine*, 5, 359–362. [1004]
LIEBMAN, J. B., AND N. MAHONEY (2017): "Do Expiring Budgets Lead to Wasteful Year-End Spending? Evidence From Federal Procurement," *American Economic Review*, 107, 3510–3549. [999]
MAHER, K. (2007): "Wal-Mart Seeks Flexibility in Worker Shifts," *The Wall Street Journal*. [997]
MARSCHAK, J., AND R. RADNER (1972): *Economic Theory of Teams*. New Haven, CT: Yale University Press. [997]
MAS, A., AND E. MORETTI (2009): "Peers at Work," *American Economic Review*, 99, 112–145. [999]
MAS, A., AND A. PALLAIS (2016): "Valuing Alternative Work Arrangements," Tech. Rep. w22708, National Bureau of Economic Research, Cambridge, MA. DOI:10.3386/w22708. [999,1016]
MESSERLI, F. H., A. W. MESSERLI, AND T. F. LÜSCHER (2005): "Eisenhower's Billion-Dollar Heart Attack—50 Years Later," *New England Journal of Medicine*, 353, 1205–1207. [1000]
MILGROM, P., AND J. ROBERTS (1988): "An Economic Approach to Influence Activities in Organizations," *The American Journal of Sociology*, 94, 154–179. [999]
MOEN, P., E. L. KELLY, W. FAN, S.-R. LEE, D. ALMEIDA, E. E. KOSSEK, AND O. M. BUXTON (2016): "Does a Flexibility/Support Organizational Initiative Improve High-Tech Employees' Well-Being? Evidence From the Work, Family, and Health Network," *American Sociological Review*, 81, 134–164. [999]
OYER, P. (1998): "Fiscal Year Ends and Nonlinear Incentive Contracts: The Effect on Business Seasonality," *The Quarterly Journal of Economics*, 113, 149–185. [999]
PERDIKAKI, O., S. KESAVAN, AND J. M. SWAMINATHAN (2012): "Effect of Traffic on Sales and Conversion Rates of Retail Stores," *Manufacturing and Service Operations Management*, 14, 145–162. [997]
POLANYI, M. (1966): *The Tacit Dimension*. New York: Doubleday Press. [1027]
PRENDERGAST, C. (2007): "The Motivation and Bias of Bureaucrats," *American Economic Review*, 97, 180–196. [999]
PRESSER, H. B. (2003): *Working in a 24/7 Economy: Challenges for American Families*. Russell Sage Foundation. [998,1000]
RADNER, R. (1993): "The Organization of Decentralized Information Processing," *Econometrica*, 61, 1109–1146. [997]

SCHUUR, J. D., AND A. K. VENKATESH (2012): "The Growing Role of Emergency Departments in Hospital Admissions," *New England Journal of Medicine*, 367, 391–393. [1004]

SHAPIRO, C., AND J. E. STIGLITZ (1984): "Equilibrium Unemployment as a Worker Discipline Device," *American Economic Review*, 74, 433–444. [999]

SIMON, H. (1947): *Administrative Behavior*. New York: Macmillan. [999]

SIMPSON, R. (1998): "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*, 9, 37–50. [1025]

SU, C.-L., AND K. L. JUDD (2012): "Constrained Optimization Approaches to Estimation of Structural Models," *Econometrica*, 80, 2213–2230. [1021]

TANABE, P., R. GIMBEL, P. R. YARNOLD, D. N. KYRIACOU, AND J. G. ADAMS (2004): "Reliability and Validity of Scores on the Emergency Severity Index Version 3," *Academic Emergency Medicine*, 11, 59–65. [1004,1006, 1007]

TIROLE, J. (1986): "Hierarchies and Bureaucracies: On the Role of Collusion in Organizations," *Journal of Law, Economics and Organization*, 2, 181–214. [999]

WEITZMAN, M. L. (1974): "Prices vs. Quantities," *The Review of Economic Studies*, 41, 477–491. [1027]

Co-editor Daron Acemoglu handled this manuscript.

Manuscript received 24 June, 2015; final version accepted 21 November, 2017; available online 17 January, 2018.