Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

The Diseconomies of Queue Pooling: An Empirical Investigation of Emergency Department Length of Stay

Hummy Song, Anita L. Tucker, Karen L. Murrell

To cite this article:
Hummy Song, Anita L. Tucker, Karen L. Murrell (2015) The Diseconomies of Queue Pooling: An Empirical Investigation of Emergency Department Length of Stay. Management Science 61(12):3032-3053. http://dx.doi.org/10.1287/mnsc.2014.2118

MANAGEMENT SCIENCE
Vol. 61, No. 12, December 2015, pp. 3032–3053
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2014.2118

The Diseconomies of Queue Pooling: An Empirical Investigation of Emergency Department Length of Stay

Hummy Song
Harvard University, Boston, Massachusetts 02163, hsong@hbs.edu

Anita L. Tucker
Brandeis University International Business School, Waltham, Massachusetts 02453, atucker@brandeis.edu

Karen L. Murrell
Kaiser Permanente South Sacramento Medical Center, Sacramento, California 95823, karen.l.murrell@kp.org

We conduct an empirical investigation of the impact of queue management on patients' average wait time and length of stay (LOS). Using an emergency department's (ED) patient-level data from 2007 to 2010, we find that patients' average wait time and LOS are longer when physicians are assigned patients under a pooled queuing system with a fairness constraint compared to a dedicated queuing system with the same fairness constraint. Using a difference-in-differences approach, we find the dedicated queuing system is associated with a 17% decrease in average LOS and a 9% decrease in average wait time relative to the control group—a 39-minute reduction in LOS and a four-minute reduction in wait time for an average patient of medium severity in this ED. Interviews and observations of physicians suggest that the improved performance stems from the physicians' increased ownership over patients and resources that is afforded by a dedicated queuing system, which enables physicians to more actively manage the flow of patients into and out of ED beds. Our findings suggest that the benefits from improved flow management in a dedicated queuing system can be large enough to overcome the longer wait time predicted to arise from nonpooled queues. We conduct additional analyses to rule out alternate explanations for the reduced average wait time and LOS in the dedicated system, such as stinting and decreased quality of care. Our paper has implications for healthcare organizations and others seeking to reduce patient wait time and LOS without increasing costs.

Keywords: pooling; fairness; queue management; strategic servers; empirical operations; healthcare
History: Received July 23, 2013; accepted October 23, 2014, by Serguei Netessine, operations management.
Published online in *Articles in Advance* May 28, 2015.

1. Introduction

Improving efficiency and customer experience is a key objective for service organizations. Skillful application of operations management principles may help achieve these goals. In particular, queue management decisions—such as queue structure and job routing policies—may impact how long customers have to wait for service and their service times.

Prior work has demonstrated through analytical models that pooling separate streams of identical customers into a single queue served by a bank of identical servers is more efficient than having a set of dedicated queues, because pooling results in shorter wait times for service (Eppen 1979, Kleinrock 1976). Having a pooled queue structure leads to a reduction in wait time because it enables customers to be processed by any available server from a bank of servers, rather than having to wait for a specific server to become available. That said, prior analytical research also suggests that pooling queues may not always yield the expected performance improvements (Debo et al. 2008, van Dijk and van der Sluis 2009, Hopp et al. 2007, Jouini et al. 2008, Loch 1998, Mandelbaum and Reiman 1998). For example, combining streams of customers who have different processing requirements can introduce inefficiencies that erode the benefits of pooling (Benjaafar 1995, Green and Nguyen 2001, Mandelbaum and Reiman 1998, Rothkopf and Rech 1987). In addition, the perceived unfairness of a pooled queue, in which faster servers are assigned more customers than their peers, may negatively impact the speed at which servers work (Doroudi et al. 2011). Thus, the overall impact of queue pooling in service settings is ambiguous.

To our knowledge, there have been few field-based, empirical studies on the impact of pooled versus dedicated queue management systems on the speed of service. This is an important omission because, in service settings, servers can adjust how they manage their work to increase or decrease their service rate (Doroudi et al. 2011, Hopp et al. 2009). Operations management scholars advocate for more studies that

examine how human behavior can alter the dynamics between operational variables and performance (Boudreau et al. 2003, Jouini et al. 2008). Thus, empirical research that examines the impact of queue structure on servers' behaviors can provide new insights for operations management theory and increase the relevance of queuing theory and research to practice.

To address this gap, we leverage the introduction of a new policy that changed the queuing system in only one part of a hospital's emergency department (ED), but not the other, from a pooled system to a dedicated system. The parallel trend in performance of the two parts of the ED before the queuing system change, and the fact that the change affected only one part of the ED, allows us to use a difference-in-differences approach to empirically test the impact of a change in the structure of the queuing system on the average wait time to be seen by an ED physician and the average length of stay (LOS) in the ED. LOS is a measure of service time and starts with the time the physician begins delivering care to the patient and ends with either a bed request for admission to the hospital or the discharge of a patient to their home or an outside facility. We use the term LOS rather than service time to more clearly convey that this measure encompasses both (a) the value-added time when clinicians are providing care, as well as (b) the time that the patient is occupying an ED bed but is not receiving active care (e.g., when the physician is waiting for test results or treating other patients).

The ED under study switched from a pooled to a dedicated queuing system to be able to handle the larger volume of patients predicted to occur because of the closing of a nearby ED. For both the pooled and dedicated queuing systems, a fairness constraint in the form of a round robin (RR) routing policy was used to assign patients to physicians, in which patients were evenly distributed across physicians independent of physician speed or idle time. The ED had this policy because physicians were paid a fixed salary and did not receive additional compensation for treating more patients or working more hours than scheduled. As a result, there were few financial incentives available to increase physician productivity, and instead, work was allocated equally among physicians. Using a difference-in-differences approach, we find that, on average, the use of a dedicated queuing system with a RR routing policy as a fairness constraint—after controlling for individual patient, physician, time, and ED characteristics—is associated with a 17% decrease in patients' average LOS and a 9% decrease in their average wait time relative to the control group. This represents a 39-minute reduction in LOS and a four-minute reduction in wait time—a meaningful time savings for the ED.

Operations management theory suggests a possible reason why the pooled queuing system with a fairness constraint is associated with a longer average LOS than the dedicated queuing system with a fairness constraint. Similar to workers in other service settings (Debo et al. 2008, Hasija et al. 2010, Tan and Netessine 2014), physicians in the dedicated queuing system are strategic servers who change their behaviors in response to their assigned responsibilities and ownership over the work routines and resources needed to accomplish those responsibilities (Cachon and Zhang 2007; Gilbert and Weng 1998; Hopp et al. 2007, 2009). Interviews with physicians suggest that, in this context, the increased ownership that stems from a dedicated queuing system with a fairness constraint leads to a situation in which the improvements in service rates due to better flow management are greater than the variability-buffering benefits of a pooled queuing system with a fairness constraint.

This paper makes a contribution to the literature on queue pooling because prior research has emphasized customer behaviors that reduce the process losses of dedicated queues, but fewer papers have empirically examined the impact of employee behaviors on the performance of dedicated versus pooled queuing systems (Boudreau et al. 2003, Hopp et al. 2007, Jouini et al. 2008). Our work thus informs the debate over the benefit of a pooled queue, which enables flexibility in the routing of jobs to servers, and a dedicated queue, which enables improvement in wait times and service times through better flow management.

2. Prior Research and Hypotheses

2.1. Prior Research on Queue Management and Service Times

Operations scholars have investigated at least two different contexts in which pooling may occur: inventory waiting to be processed (production-inventory systems) and customers waiting for service (queuing networks). Most closely related to our research context, studies of queuing networks focus on the effect of pooling queues of customers, servers, and tasks in service organizations (Mandelbaum and Reiman 1998). Much of this research has been conducted with call centers, and has shown that the benefits of flexible servers and pooled queues can outweigh potential drawbacks (Anupindi et al. 2005, Bassamboo et al. 2010, Gans et al. 2003, Jouini et al. 2008). Researchers have reached similar conclusions in other settings, such as mail delivery, finding that pooling can improve quality while concurrently reducing costs (Ata and Van Mieghem 2008). Furthermore, prior research has found that pooling is beneficial and wait time reductions are achieved even when work

is allocated fairly among servers using a RR routing policy (Hyytiä and Aalto 2013, Raz et al. 2006). In fact, Armony and Ward (2010) find that pooling with a fairness constraint outperforms classical pooling when the arrival rate of customers is high because faster servers have an incentive to slow their service rate under systems in which work is allocated based on server availability instead of a fair distribution across servers.

On the other hand, some analytical models have shown that the behavioral responses of servers and customers can reduce the expected benefits of queue pooling (van Dijk and van der Sluis 2008, Hopp et al. 2007, Loch 1998, Mandelbaum and Reiman 1998, Rothkopf and Rech 1987). Most pertinent to our study, strategic servers may reduce the effectiveness of queue pooling (Cachon and Zhang 2007; Debo et al. 2008; Hopp et al. 2007, 2009; Jouini et al. 2008). First, they may manipulate customer service times to be higher or lower by managing their tasks differently when it benefits them to do so (Hopp et al. 2007, Link and Naveh 2006, Tan and Netessine 2014). For example, in the restaurant industry, Tan and Netessine (2014) find that wait staff adjust the services offered to customers so that customers spend less time in the restaurant when their workload is high. Similarly, Oliva and Sterman (2001) find that bank employees reduce the steps they go through to approve loans when their workload is high, even though this erodes bank profitability. Second, strategic servers can also slow down their work pace. Using analytical models, Debo et al. (2008) show that when workers are paid by the quantity of work completed, such as taxicab drivers and lawyers, they add unnecessary tasks when business is slow, thereby increasing service time for their customers. Similarly, Hasija et al. (2010) find that call center agents take more time to answer customers' queries when they have low workloads if their contract rewards them for keeping utilization above a minimum threshold. Collectively, these studies suggest that service time is impacted by strategic servers' responses to incentives and responsibilities.

Even when strategic servers do not have direct financial incentives to adjust their service rates, they may still manipulate their service times if they have a high degree of perceived ownership over their assigned jobs. Employees feel higher levels of ownership when they are given the resources and responsibility to manage the complete workflow of a meaningful task (Hackman and Oldham 1976). By design, dedicated queuing systems with a fairness constraint afford higher levels of ownership than do pooled queuing systems with the same fairness constraint because in the former, each server has been explicitly assigned the responsibility for efficiently completing the work waiting in his or her queue. In contrast, pooled queuing systems provide lower levels of ownership because the responsibility for depleting the queue is dispersed over multiple servers. Thus, strategic servers in dedicated queuing systems with a fairness constraint may be more motivated to efficiently manage their workload than those in pooled queuing systems with a fairness constraint (Doroudi et al. 2011, Gilbert and Weng 1998).

2.2. Queue Management and Strategic Physician Behavior in the Emergency Department

ED physicians are strategic servers, as defined by Cachon and Zhang (2007). To illustrate how physicians operate as strategic servers, consider an ED physician who has a patient with a headache. The physician can treat the patient using any combination of the following tasks: obtain a detailed medical history to generate possible causes of the headache, order a computed tomography scan, or prescribe aspirin. The physician's choice can impact the patient's LOS because of the variance in time required for the different options. In addition, the physician can influence patient LOS by proactively pulling for information, such as x-ray results, rather than waiting for that information to be pushed. The physician can also control his or her own utilization because there are usually multiple patients under the care of an ED physician. Thus, physicians can reduce their own idle times and further increase the flow of patients through the system.

In this paper, we consider two different types of queuing systems in the context of an ED. In a pooled queuing system—which is typical for most EDs in the United States—a physician is assigned to a patient only once the patient is placed in an ED bed. This means patients in the waiting room remain in a pooled queue while waiting for an open bed. In a dedicated queuing system, physicians are assigned to patients at the point of triage. Here, patients in the waiting room are, in effect, waiting to be seen by a specific physician. In the dedicated queuing system, each physician thus has a greater ownership over his or her workload even before the patient is placed in an ED bed.

In the ED that we study, each physician in the dedicated system also controls his or her own bank of resources (e.g., beds and nurses) necessary to facilitate the flow of his or her own patients. Physicians are assigned patients in a RR fashion that fairly allocates patients among all physicians independent of physicians' service rates. In addition, they can only go home when all of their assigned patients are discharged or nearly discharged (e.g., awaiting a test result), and are not paid extra for working past the scheduled end of their shift. Therefore, physicians have an incentive and the ability to manage

their workload as efficiently as possible. For example, physicians can coordinate the care of their patients with their nurses to prioritize getting test results back for a patient so he can be discharged, and then quickly move a patient from the waiting room into that vacated bed. In contrast, in the pooled queuing system, physicians do not "own" patients in the waiting room, nurses and beds are shared among all physicians, and they rely on a triage nurse, called the "internal triage" nurse, to manage the flow of patients into available beds for the entire ED. Thus, in the pooled queuing system, physicians' have ownership over a much smaller portion of the patient flow process. Based on our interviews with physicians and observations of their practice patterns, we suspect that the higher level of ownership of one's workload and the resources necessary to manage that workload afforded by the dedicated queuing system increases physicians' perceived ownership over patient flow. This results in physicians having a faster rate of discharging patients throughout their entire shift than when in the pooled system.

Prior theoretical operations management research suggests that when strategic servers have ownership and responsibility for managing flow, it can lead to lower service times. Gilbert and Weng (1998) and Cachon and Zhang (2007) construct analytical models of a buyer's choice of queue structure for allocating demand among two suppliers. They find that suppliers in a dedicated system produce the goods faster than those in a pooled system because the dedicated system's suppliers have more incentive to invest in production capacity. The dedicated system provides certainty that they will benefit from their capacity investments, which can be thought of as having ownership over a demand stream in combination with the responsibility over production resources needed to meet that demand. Similarly, in the context of a hospital's inpatient department, Best et al. (2015) use a stylized model to show that a patient flow director with increased ownership and responsibility for managing flow is able to attain a significant decrease in patient LOS. The authors suggest that this decrease is attained from increased motivation to cut non-value-added time and better coordinate patient care among doctors, nurses, and case managers.

In the context of an ED, switching from a pooled to a dedicated queuing system should similarly affect the behavior of physicians by increasing the degree of ownership physicians have over their patients' flow through the ED. Specifically, we hypothesize that ED physicians may attain a shorter average LOS for their patients when they work in an ED with a dedicated queuing system with a fairness constraint. Prior research suggests that servers work slower at low workloads because there is no need to work fast because of the slack capacity (Tan and Netessine 2014). However, in our ED setting, workloads are typically at high levels because of the ED's ability to staff according to historical demand and to send clinicians home early during periods of unexpectedly low demand. Therefore, we hypothesize a direct, positive effect of a dedicated queuing system on LOS.

Hypothesis 1. *LOS is shorter in the ED when physicians are working in a dedicated queuing system as opposed to a pooled queuing system.*

We further consider how dedicated queuing systems may affect patients' average wait times. A priori, it is unclear whether dedicated queues with strategic servers will result in shorter or longer wait times for customers. On one hand, when under a dedicated queuing system, if patients currently being cared for spend less time in an ED bed and if a physician proactively places the next patient from his or her queue into the newly available bed, the next patient's wait time may decrease because of an indirect queuing effect. In other words, the benefits of a dedicated queue—fair assignment of work and ownership over patients, resources, and patient flow—may overcome the negative impact on wait time of using a dedicated rather than a pooled queue. Thus, we predict the following:

Hypothesis 2A. *Wait time is shorter in the ED when physicians are working in a dedicated queuing system as opposed to a pooled queuing system.*

On the other hand, switching from a pooled to a dedicated queue may result in an increase in wait time, due to the well-known inefficiency of forcing customers' whose server is busy to wait for that server to be free, even if another server is idle (Eppen 1979, Kleinrock 1976). The inefficiency of dedicated queues might overpower the possible reduction in wait times because of faster service times. Therefore, we test the following competing hypothesis.

Hypothesis 2B. *Wait time is longer in the ED when physicians are working in a dedicated queuing system as opposed to a pooled queuing system.*

To understand the behavioral mechanism through which different queuing systems may impact LOS, we explore the rates at which physicians discharge patients during different time periods throughout their shifts. We hypothesize that the higher level of ownership over patient flow afforded by a dedicated queuing system, as opposed to a pooled queuing system, motivates physicians to more efficiently manage patient flow throughout the duration of the entire shift. Physicians in the dedicated system may be able to efficiently manage patient flow—and thus achieve higher discharge rates—by proactively "pulling" for

lab, x-ray, and specialty consult results; improving coordination with nurses to prioritize tasks necessary for discharge; initiating the discharge process sooner for patients ready for discharge; and making sure that nurses place waiting patients into available beds as soon as possible. This hypothesized increase in discharge rate is in contrast to only speeding up toward the end of the shift, which would be predicted if physicians were only subject to a deadline effect and were not better managing patient flow (Deo et al. 2014).

Prior theoretical research suggests that physicians in the dedicated system will have a greater incentive to consistently work at a higher rate because they can reap the benefits that stem from achieving a faster rate of production (Gilbert and Weng 1998). In our setting, the benefits to physicians of obtaining a higher discharge rate are (a) more time to spend with current patients, which increases both patient and physician satisfaction (Hopp et al. 2007); (b) idle time if the physician has no additional patients currently in queue (Armony and Ward 2010); and (c) less work remaining for the physician to complete before he or she can go home. In a pooled queuing system, these benefits do not necessarily accrue to physicians who work at a higher rate because the misalignment of responsibility for patient flow and ownership over patients and resources prevents physicians from being able to reap these benefits. Thus, we hypothesize that physicians working in a dedicated queuing system will attain higher rates of discharging patients throughout the shift. Specifically, we hypothesize that this increase in discharge rate will emerge a few hours after the beginning of a shift because the average LOS is greater than two hours and, therefore, it would not be possible to discharge many patients in the first two hours of one's shift. However, after this initial two-hour period, the faster discharge rate will be present throughout the remainder of the shift, rather than only at the end of the shift.

Hypothesis 3. *A physician's discharge rate of patients is greater for each noninitial time period of the shift when physicians are working in a dedicated queuing system as opposed to a pooled queuing system.*

3. Setting, Data, and Empirical Methods

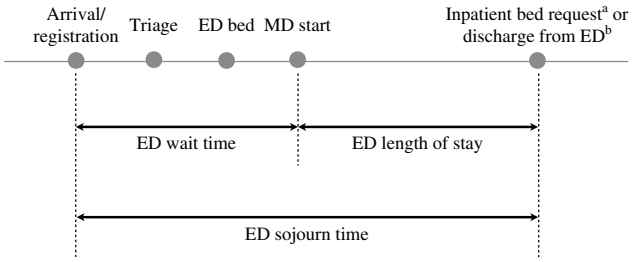
3.1. Research Setting

Our data comes from the ED of a 162-bed hospital in Northern California. We select this ED for study because in August 2008, it experienced an intervention—which we describe in more detail in §3.2—that transformed a part of the ED from having a pooled queuing system to a dedicated queuing system for the patients waiting to be seen in the ED. We use data from a timespan before and after the intervention (March 2007 to July 2010) to test our hypotheses about the impact of queuing systems on average LOS, wait time, and discharge rate in the ED.

Depending on the time of day, this ED had an average of two to five physicians staffing 41 ED beds and up to nine hallway gurneys. One bed was located in the resuscitation room and reserved for patients arriving without a pulse, three beds were in the trauma bay reserved for trauma intakes, four beds were in the rapid care area (RCA) for low severity patients, and a minimum of two beds were reserved for psychiatric patients. This ED experienced an average 5% increase in patient volume each year, from approximately 65,000 patients in 2007 to 76,000 patients in 2010. The average daily patient volume was 178 patients in 2007 and 212 patients in 2010. This was a relatively large patient volume in comparison to other EDs in the surrounding areas.

This ED, like many others, had a standardized patient flow process (Figure 1). On a patient's arrival, a registration clerk conducted a brief registration process. A second triage nurse, called the "external triage" nurse, obtained vital signs, collected the chief complaint, and assigned an Emergency Severity Index (ESI) triage category—a commonly used, standard ranking of ED patient severity that ranges from levels 1 (highest acuteness) through 5 (lowest acuteness). This triage process accounted for a patient's expected level and type of resource utilization, and was used to route a patient to either the main area (main ED) or the RCA. The two areas of the ED each had its own equipment and staff to deliver care to patients (e.g., the RCA had its own computer terminals and vital sign monitors that were separate from the main ED's equipment). Ninety-eight percent of higher acuteness patients (ESI levels 1, 2, or 3) were treated in the main ED. Seventy-five percent of lower acuteness patients (ESI levels 4 or 5) were treated in the RCA. Lower acuteness patients were treated in the main ED when main ED beds were available and the waiting room census was low (15% of lower acuteness patients) or when they arrived between 11 p.m. and 7 a.m. when the RCA was closed (9% of lower acuteness patients).

Figure 1 Standard Patient Flow in the Emergency Department



[a]For patients who were admitted to the hospital.
[b]For patients who were discharged home or to an outside facility.

In this ED, a computer system assigned each patient to a specific attending physician, either on assignment to a bed (pooled queuing system) or at the point of triage (dedicated queuing system). The assigned physician assumed responsibility for completing the set of physician-related tasks for that patient during the patient's ED visit, such as taking the patient's history, prescribing medications, and ordering tests or treatments. This physician could consult with other physicians concerning his or her patient's care, but this did not transfer the responsibility for patient care to the consulting physician. It was common for a physician to serve multiple patients simultaneously. In other words, a physician did not need to discharge one patient before starting work for the next patient.

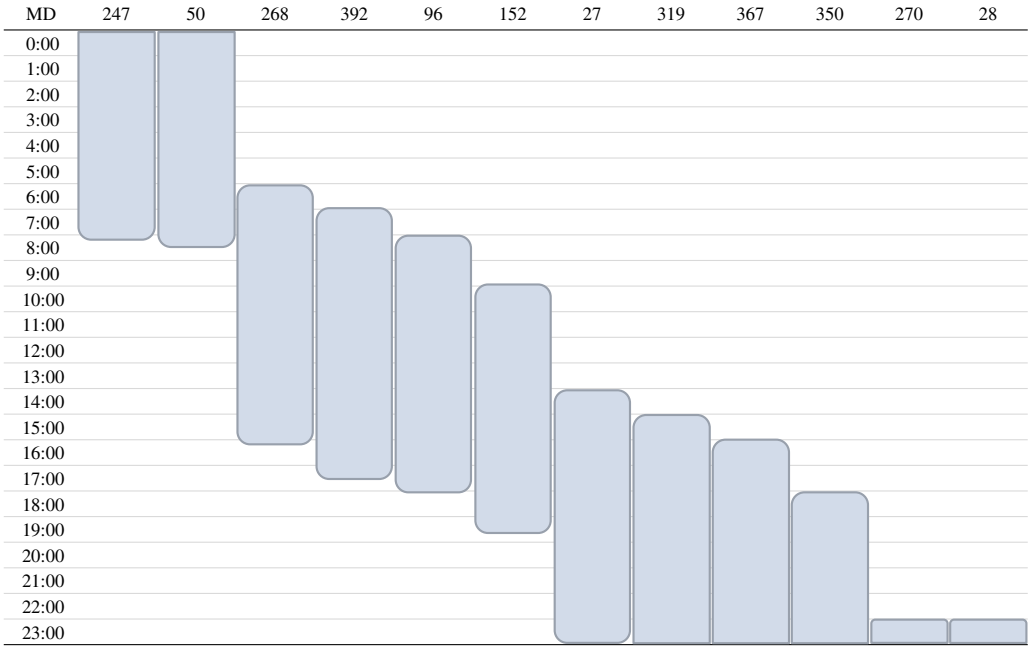
Physicians arrived at staggered times throughout the day, such that there was not a certain time at which all physicians changed shifts (Figure 2). Physician shift times were determined in advance by the ED chief, and the ED scheduler assigned individual physicians to the predetermined shift times. Physicians could change shifts on the hour between 5 a.m. and 11 a.m., between 2 p.m. and 5 p.m., and at 11 p.m. or midnight. Between 7 a.m. and 11 p.m., there was usually one physician working in the RCA and four physicians working in the main ED. During the overnight shift from 11 p.m. to 7 a.m., there were a minimum of two physicians and a maximum of four physicians working in the main ED.

Physicians were assigned to either the RCA or the main ED for the full duration of a shift by the ED scheduler. They were paid a flat rate for their shift without any additional compensation for the services provided or the number of hours worked. Thus, there were no incentives to stretch out treatment times by providing additional services. Prior to leaving the shift, physicians were expected to discharge or at least complete a care plan for the cohort of patients assigned to them (e.g., indicate what next steps should be taken if the lab test comes back positive versus negative), which incentivized physicians to get their patients through the system as efficiently as possible. Physicians were not required to stay if they had patients who were simply boarding in the ED, waiting to be transferred to an inpatient unit or to another facility. To allow physicians enough time to either complete a care plan or discharge the patients who had been assigned to them, they were assigned new patients only up until two hours before the scheduled end of their shifts. Patients arriving during the last two hours of a physician's shift were assigned to one of the other physicians on shift or, if it was close enough in time, to the oncoming physician. Because physician shifts were sufficiently staggered, there was always a physician available to take newly arriving patients and this did not induce a greater variation in system productivity.

3.2. Intervention: Change in the Patient Assignment System

In August 2008, the main ED implemented an intervention called the patient assignment system (PAS). PAS restructured the main ED from having a pooled queuing system to a dedicated queuing system. Prior

Figure 2 **(Color online) Example of Physician Shift Distribution Over a 24-Hour Period**



Notes. MD numbers across the *x* axis are unique physician identifiers. Shaded bars indicate the duration of a physician's shift.

to PAS, higher severity patients due to be seen in the main ED returned to the waiting room after being triaged, with the exception of ESI level 1 patients who proceeded directly to the resuscitation room. When a bed became available in the main ED, the internal triage nurse placed the next patient of highest severity in this bed. Our interviews with ED physicians revealed that this process often resulted in a delay from the bed becoming available to a patient being placed in the bed because the internal triage nurse was not responsible for patient flow through the ED, and the physicians did not feel responsible for making sure that empty beds were filled quickly. Once a patient was placed in a bed, the computer system assigned each patient to a physician using an RR routing policy, which means that each patient was assigned to a physician in a set order that evenly distributed patients among physicians regardless of each physician's current workload. Once this assignment occurred, the physician could see the assigned patient listed under his or her panel when logged onto the patient management system on one of the ED computers. Thus, when a patient was waiting in the waiting room, he or she was in a pooled queue waiting to be assigned to any one of the, on average, four physicians on shift in the main ED.

Prior to the PAS intervention, the patient only entered a specific physician's queue after being placed in an available main ED bed by the internal triage nurse. It was at this point that the physician had ownership of the patient, not before. The only exception to the RR routing policy was made when a physician was currently involved in the resuscitation of an ESI level 1 patient, in which case another physician could voluntarily take on that physician's next patient. In addition, at the beginning of a physician's shift, the computer system assigned one, two, or three consecutive patients to the oncoming physician. The specific number of consecutive patients to whom a physician was assigned was automatically determined by the computer system based on the rate of patient arrivals. This RR routing policy was instituted to prevent physicians from unfairly selecting "easier" patients and to ensure that the faster physicians would not be unequally assigned more work simply because of their higher service rates. This simultaneously made patient routing to physicians both fair and nearly random rather than due to a physician's seniority or speed of discharging patients. It was feasible to implement because there were two organizational structures in place to minimize the variation in workload across the physicians staffing the main ED: (a) the hospital's trauma team assumed primary responsibility for incoming trauma patients and thus did not disproportionately increase the workload of an ED physician; (b) the RCA cared for lower severity patients. Thus, there was limited variation in patient intensity among the patients being assigned to the physicians staffing the main ED.

After PAS implementation, the computer system still used the RR routing policy but assigned each patient to a physician at the point of triage. This means that, when a physician logged onto the patient management system to view his or her panel of patients, the display showed not only those patients who were already placed in ED beds but also those who were still in the waiting room. This increased physicians' perceived ownership of their patients because they were responsible for their patients' care and experience from triage onward—which included their time in the waiting room—rather than just from placement in an ED bed. In conjunction, it was now the physicians' responsibility to make sure their next patient from the waiting room was placed in an available main ED bed. To enable physicians to carry out this additional responsibility, six main ED beds and two hallway gurneys were allocated to each physician working in the main ED. In addition, two nurses were assigned to each physician to help care for patients, although each physician typically worked with other nurses outside of these two nurses during the course of the shift because (a) nurses' shift change times were not aligned with that of the physician and (b) nurses had designated break times during which a relief nurse substituted in for the duration of the break.

After PAS implementation, the computer system's RR routing policy was maintained and adhered to, even if there was a physician who had waiting patients while another physician had an available ED bed and no waiting patients. Hence, patient assignment remained independent of a physician's speed of discharging patients. Similarly, the incentive of having to stay until all patients had been cared for remained constant, though now physicians also had to care for the patients who had been assigned to them who were still in the waiting room.

In the RCA, the process used to assign patients to the physician working in the RCA did not change over the course of our study. A lower severity patient was assigned to a physician when he or she was called to be seen in the examination room, not while in the waiting room. Thus, the RCA physician was not responsible for any patient who was still waiting in the waiting room at the conclusion of his or her shift; any patient still waiting became the responsibility of the next physician coming on to the shift.

3.3. Data

This study uses approximately three and a half years of de-identified electronic medical record (EMR) data of all 238,946 patients treated in the ED from March 1, 2007 to July 31, 2010. The data set contains patient-level information including, but not limited to, the

following: the patient's time of arrival and departure, LOS, ESI level, attending physician, and disposition. We exclude patients with no attending physician or ESI level listed on their record, patients who left without being seen by a physician, patients who had a LOS of zero minutes or less, patients whose records lacked a time stamp for when the physician began caring for the patient, and patients who were admitted to the hospital but whose records lacked a time stamp for when a bed request was made. In addition, we exclude patients whose LOS was greater than 48 hours; most of these patients presented with a psychological condition and were waiting to be discharged to an appropriate facility. We exclude these observations from our data set because their extended LOS was typically driven by placement logistics rather than by physicians' levels of productivity. In addition, we exclude patients of ESI level 1 (i.e., patients needing resuscitation) and patients who died in the ED because their LOSs were likely to be driven by factors other than physician productivity. Finally, we exclude trauma patients because the hospital's trauma team, not a particular ED physician, primarily cared for these patients. Altogether, we exclude 12,817 patients or 5.4% of the overall sample.

Using this sample of 226,129 patients, we create a patient-level panel data set that treats the physician as the panel variable. For our analyses, we exclude data from August 2008 to account for an acclimation period because the exact date of PAS implementation is unknown. In addition, we limit our sample to the patients seen by physicians who were full-time employees of this ED. Physicians who worked in this ED but were not full-time employees tended to be employees of other hospitals in the hospital's network who were brought in to cover small portions of shifts when the full-time ED physicians were not able to staff the ED (e.g., during physician staff meetings). This results in a final sample of 217,213 patients.

In addition to the EMR data, we also gathered qualitative data through 86 hours of observations of ED staff and unstructured interviews about workflow in the ED with ED physicians, nursing staff, and the ED unit leadership.

3.4. Dependent Variables

Our key dependent variables are ED wait time, ED LOS, and patient discharge rate. ED wait time is defined as the time from a patient's arrival to the ED to the time the physician began delivering care. ED LOS starts with the time the physician began delivering care to the patient and—for patients admitted to the hospital—ends with a bed request for admission to the hospital, thus excluding the time spent boarding in the ED and any time spent in an inpatient unit. For patients discharged to home or to an outside facility, ED LOS ends at the time of discharge. We log-transform ED wait time and LOS because each of their distributions are otherwise right skewed. Patient discharge rate is defined as the number of patients discharged per hour by a given physician in a specified two-hour period of the shift, such as the first two hours, second two hours, or final two hours.

We employ a set of additional dependent variables for analyses that extend the main findings and consider possible alternative explanations. These include binary indicators for whether a lab was ordered, an x-ray was ordered, a patient was admitted to the hospital, a patient died in the ED, or a patient returned to the ED within 72 hours, respectively.

3.5. Independent and Control Variables

3.5.1. Patient Assignment Intervention in the Main ED. The implementation of PAS marks the time at which the main ED transitioned from having a pooled queuing system to a dedicated queuing system. We capture this transition with a binary interaction term, $PAS \times main$, which is equal to 1 in the main ED after the implementation of PAS and 0 otherwise (i.e., in the main ED before the implementation of PAS, or in the RCA at any time). To account for an acclimation period, we designate the pre-PAS period to include up to July 31, 2008 and the post-PAS period to begin with September 1, 2008.

3.5.2. Control Variables. We account for several factors that may affect our dependent variables and may be correlated with our independent variables, *PAS* and *main*. These include factors related to the patient's condition, the state of the ED, the physician's practice experience, and time trends. To account for the variation in LOS due to the severity of a patient's condition, we control for the patient's acuteness and age. We account for patient acuteness using a series of dummy variables that reflect ESI levels 2, 3, 4, and 5, respectively. The combination of a patient's ESI level and age is the best approximation we have for patient condition and severity because our data set does not include patients' specific diagnoses (e.g., diagnosis-related groups (DRGs)). It is important to control for patient acuteness because the patient mix in this ED changed over time, wherein more patients presenting to the main ED were of higher acuteness and more patients presenting to the RCA were of lower acuteness after PAS implementation.

To capture ED busyness and congestion, we control for the total number of physicians working during a given morning, afternoon, or overnight shift; the number of patients waiting to be seen by this physician at a given time; the number of patients being seen by this physician at a given time; whether an ESI level 1 patient was present in the ED; and whether a trauma

patient was present in the ED. Relatedly, to account for other systematic differences in patients' LOSs that would arise from differences in structural elements of the ED, we control for the general time frame of the physician's shift (morning, afternoon, or overnight) and the location of the shift (main ED or RCA).

To account for systematic differences arising from differences in physicians' experience working in this particular ED, we control for the number of shifts the physician has worked in this ED since the beginning of the data set up until the point of each patient encounter. As we explain in more detail in §3.6, we also include physician fixed effects to account for other unobserved differences by physician.

Finally, we account for time trends and related influences by including dummy variables for day of the week and by using month-year fixed effects.

3.6. Empirical Models

Our main analyses use a difference-in-differences framework to examine the relative changes in LOS and wait time for patients seen in the main ED and the RCA before and after PAS implementation. We use linear regression models with month-year and physician fixed effects and clustered standard errors. We cluster standard errors by physician to account for within-physician correlations of the error terms, both within and across shifts, rather than imposing the usual assumption that all error terms are independently and identically distributed. The fixed-effects models allow us to capture time trends and to control for unobservable individual physician effects that do not vary over time, such as level of motivation, innate ability, and practice routines. These are important to account for because they may significantly influence a physician's productivity level in ways that cannot be measured (McCarthy et al. 2012).

In addition to the standard assumptions of linear regression models, fixed-effects models make two key assumptions, both of which are satisfied in our study. First, is the assumption of strict exogeneity, which means the observation-specific error term is uncorrelated with the covariates of the observation and all other observations belonging to the same cluster (Wooldridge 2010). This is a plausible assumption in our context because (a) there is a low likelihood that patients with multiple visits are treated by the same physician and (b) the patient error term is unlikely to be correlated with the covariates for other patients of the same physician. In addition, the unobservable random traits of physicians that affect their patients' average LOS are not likely to be associated with the key independent variable of interest. Specifically, the RR routing policy makes it unlikely that the fastest physicians receive the most complicated cases since patient assignment to physicians is random and is not driven by physician speed or physician preference.

We use fixed-effects models rather than random-effects models because we do not believe that the random-effects assumption of zero correlation between the month-year effect or physician effect and the other covariates (such as the number of shifts worked by the physician) holds. By using fixed-effects models, we can account for the unobserved traits of each month-year and of each physician that are associated with a patient's LOS and also correlated with the independent variables of interest. Accordingly, we conduct the Durbin-Wu-Hausman test, which rejects the random-effects model in favor of the fixed-effects model ($\chi^2 > 169.45$, $p < 0.001$).

3.6.1. ED LOS. To test Hypothesis 1, we estimate the following difference-in-differences model at the patient level:

$$\ln LOS_{ijt} = \alpha_0 + \alpha_1 main_{ij} + \alpha_2 PAS_t \times main_{ij} + \delta \mathbf{X}_{ijt} + \theta_t + \gamma \mathrm{MD}_i + \varepsilon_{ijt}. \quad (1)$$

Here, $\ln LOS_{ijt}$ represents the logged number of minutes that patient i of physician j stayed in the ED in month-year t; $main_{ij}$ indicates whether patient i of physician j was seen in the main ED; $PAS_t \times main_{ij}$ is an interaction term equal to 1 when the patient was seen in the main ED after the implementation of PAS; $\mathbf{X}_{ijt}$ is a vector of patient, physician, and day-of-week covariates; θ_t is a vector of month-year fixed effects, MD_i is a vector of physician indicators; α's and δ's represent vectors of coefficients; γ represents a vector of physician fixed effects; and ε is the time-varying error term not already captured. Table 1 provides summary definitions for all variables included in our models.

In estimating Equation (1), we use the difference-in-differences estimator, $PAS_t \times main_{ij}$, to compare the difference in patients' average LOS in the main ED and the RCA before PAS implementation to the difference after PAS implementation. Because the queue structure did not change in the RCA, whereas the main ED moved from having a pooled to a dedicated queuing system, we consider the shifts worked in the RCA as comprising the untreated comparison group and those worked in the main ED as comprising the treatment group. By using a difference-in-differences approach, we are able to control for any bias caused by variables common to the main ED and the RCA, even when those variables are unobserved. Although the acuteness of patients seen in the two parts of the ED differed, thus implying differences in treatment processes and levels of patient LOS, the RCA serves as a reasonable control because, as our interviews with ED leadership and staff indicate, there were no changes besides PAS during the study period that affected only one part of the ED and not the other. Furthermore, we

Table 1 Summary Definition of Variables

Variable	Description	Level of analysis
	Main dependent variable	
ED wait time	Logged number of minutes elapsed between patient arrival to ED and MD start	Patient
ED length of stay	Logged number of minutes elapsed between MD start and bed request (for patients admitted to hospital) or discharge from ED (for patients discharged home or to an outside facility)	Patient
Discharge rate	Number of patients discharged per hour by a given physician in a given two-hour period of the shift (e.g., penultimate two hours, final two hours)	Physician-shift (two-hour period)
	Independent and control variables	
ESI level	Four indicators for patient's ESI level (from highest to lowest: 2, 3, 4, 5)[a]	Patient
Age	Patient age in years	Patient
MDs on shift	Number of all physicians working at any point during this shift	Physician-shift
Current waiting count	Number of patients waiting to be seen by this physician at this time	Patient
Current patient count	Number of patients being seen by this physician at this time	Patient
Shift number	Indicator for what number shift this is for this physician in this data set	Physician-shift
ESI level 1 *patient present*	Indicator for presence of ESI level 1 patient (= 1 for present, = 0 for absent)	Patient
Trauma patient present	Indicator for presence of trauma patient (= 1 for present, = 0 for absent)	Patient
Arrival shift type	Three indicators for type of shift during which patient arrived (morning, afternoon, overnight)	Patient
Months since March 2007	Indicator for what number month this is in this data set	Patient
Day of week	Seven indicators for day of week of shift	Patient
Main ED	Shift location (= 1 for main ED, = 0 for rapid care area)	Physician-shift
PAS implemented	Indicator for whether PAS was implemented (= 1 for preimplementation, = 0 for postimplementation)	Physician-shift
Interaction	*PAS* × *Main ED*	Physician-shift
	Additional dependent variables	
Lab ordered	Indicator for whether lab was ordered (= 1 for ordered, = 0 for not ordered)	Patient
x-ray ordered	Indicator for whether x-ray was ordered (= 1 for ordered, = 0 for not ordered)	Patient
Admitted to hospital	Indicator for whether patient was admitted to hospital on discharge from ED (= 1 for admitted, = 0 for not admitted)	Patient
Died in ED	Indicator for whether patient died in ED (= 1 for died in ED, = 0 for did not die in ED)	Patient
Revisit within 72 *hours*	Indicator for whether patient returned to ED within 72 hours (= 1 for returned, = 0 for did not return)	Patient
Shift duration	Number of hours for which physician worked in ED during this shift	Physician-shift
ED sojourn time	Logged number of minutes elapsed between arrival to ED and bed request (for patients admitted to hospital) or discharge from ED (for patients discharged home or to an outside facility)	Patient
ED boarding time	Logged number of minutes elapsed between bed request and discharge from ED (if admitted to hospital)	Patient

[a]Although the ESI uses five categories, we have four indicators for patient ESI level because we exclude patients of ESI level 1 from our analysis.

find that average LOS in the main ED and the RCA, respectively, exhibit parallel trends in the 17 months preceding the implementation of PAS.

After establishing the parallel trend assumption (Abadie 2005, Duflo 2001), we estimate the effect of transitioning from a pooled to a dedicated queuing system on patients' average LOS by examining the coefficient on the interaction term, $PAS_t \times main_{ij}$. We predict that this coefficient, α_2, is negative and statistically significant, suggesting that the dedicated queuing system is associated with a shorter average LOS than the pooled queuing system.

3.6.2. ED Wait Time. To test Hypotheses 2A and 2B, we estimate the following difference-in-differences model at the patient level:

$$\ln wait_{ijt} = \beta_0 + \beta_1 main_{ij} + \beta_2 PAS_t \times main_{ij} + \delta \mathbf{X}_{ijt} + \theta_t + \gamma \mathrm{MD}_i + \varepsilon_{ijt}. \quad (2)$$

In this model, all variables remain the same as in Equation (1) with the exception of $\ln wait_{ijt}$, which represents the logged number of minutes that patient i of physician j in month-year t spent in the waiting room on arrival to the ED. We use the same difference-in-differences approach as we do in testing Hypothesis 1. Here, we estimate the effect of PAS on patients' average ED wait time by examining the coefficient on the difference-in-differences estimator, $PAS_t \times main_{ij}$. Hypothesis 2A predicts that this coefficient, β_2, is negative and statistically significant because of an indirect queuing effect, suggesting that the dedicated queuing system is associated with a shorter average wait time than the pooled queuing system. Hypothesis 2B predicts that β_2 is positive and statistically significant because of the inefficiency of dedicated queues, suggesting that the dedicated queuing system is associated with a longer average wait time than the pooled queuing system.

3.6.3. Discharge Rate. To test Hypothesis 3, we estimate the following model at the physician-shift two-hour period level:

$$\ln dischrate_{kj} = \varphi_0 + \varphi_1 PAS + \delta \mathbf{X}_{kj} + \gamma \mathrm{MD}_k + \varepsilon_{kj}. \quad (3)$$

Here, $dischrate_{kj}$ represents the number of patients discharged per hour by physician j in a given two-hour shift period k; *PAS* indicates whether PAS had been implemented; φ's and δ's represent vectors of coefficients; and all other variables remain the same. For this analysis, we limit the sample to patients seen in the main ED and conduct a pre-post analysis. We do not employ a difference-in-differences approach because the different discharge processes in the main ED and the RCA make the comparison difficult, and because we are interested in the change in discharge rates during each of the two-hour periods over the course of a physician's shift rather than the change in the average discharge rate of a physician's shift. Therefore, we estimate Equation (3) separately for the first, second, penultimate, and final two-hour periods of a physician's shift. This allows us to examine whether and at what point during a physician's shift the implementation of PAS in the main ED affects the discharge rate of patients. Because the discharge rate is small and discrete and because the data are not overdispersed, we employ a Poisson model with physician fixed effects.

If the dedicated queuing system results in a reduction in patients' average LOS, we would expect the discharge rate in each of the two-hour periods of a physician's main ED shift to increase after PAS implementation. This is because, after PAS implementation, physicians are more likely to engage in strategic behaviors throughout the shift to ensure that their patients' average LOS is as short as possible. However, because many of the preliminary tasks may be unaffected by the post-PAS increase in ownership, we expect that the discharge rate may be unaffected in the first two-hour period of a physician's shift. Accordingly, we predict that the coefficient on *PAS* will be positive and statistically significant for each of the second, penultimate, and final two-hour periods of a physician's shift, whereas it will not exhibit a statistically significant change for the first two-hour period of a shift.

3.6.4. Additional Analyses. To better understand our main findings and consider possible alternate explanations, we conduct several additional analyses. We begin by considering two competing explanations that could account for the decrease in average LOS post-PAS. First, patients might have experienced shorter LOSs in the ED because physicians "cut corners" by stinting on care (Oliva and Sterman 2001). We assess this possibility by estimating Equation (1) with two different dependent variables, both measured at the patient level: whether labs are ordered for a patient and whether x-rays are ordered for a patient. Data on whether labs or x-rays are ordered for a patient are obtained directly from the hospital's EMR system. For each of these variables, we estimate Equation (1) as a logistic regression because both are binary indicator variables. Second, we consider whether the decrease in LOS stems from physicians shifting their work onto other clinicians. In the context of the ED, the most plausible scenario is ED physicians admitting more patients to the hospital, so that patients appear to stay in the ED for a shorter period of time. We examine this possibility by estimating Equation (1) with admission to the hospital as the dependent variable. Data for whether the patient is admitted to the hospital comes from the EMR system and is measured at the patient level. Again, we estimate a logistic regression because admission to the hospital is a binary dependent variable. Next, we examine the possibility of the quality of care in the ED declining as an unintended consequence of PAS implementation in the main ED. As proxies for quality, we examine whether the patient returned to the ED within 72 hours after an initial visit and whether the patient died in the ED. We estimate Equation (1) as a logistic regression with each of these binary indicators as the dependent variable, respectively. For the analysis of ED revisits, we employ a 72-hour time period, which is the standard quality metric used to capture returning ED patients (Keith et al. 1989). For the analysis of patient mortality in the ED, we include a subset of previously excluded patient-level observations—specifically patients of ESI level 1, patients who died in the ED, and trauma patients.

In addition, we consider the potential impact of PAS on the duration of a physician's shift, which is measured as the number of hours for which a physician worked in the ED during a particular shift. Though this does not directly address why having a dedicated queuing system may decrease patients' average LOS, it is an important consideration if implementing a similar system at other EDs. If having a dedicated queuing system results in physicians staying longer to finish caring for their assigned patients, it may not be feasible to implement elsewhere for reasons of cost and physician burnout. To assess this possibility, we estimate a regression of a similar form as Equation (1) but with the shift duration as the dependent variable and at the physician-shift level. We use the shift duration and not the log of shift duration because the variable is normally distributed. We estimate this regression at the physician-shift level because the dependent variable (i.e., shift duration) is calculated at this level. If physicians are working longer hours as a result of PAS implementation, we would expect to see a positive and statistically significant coefficient on the interaction term, $PAS \times main$.

Finally, we examine the impact of PAS implementation on sojourn time, which is the sum of ED wait time and LOS. We also examine the impact of the queue structure on ED boarding time to assess whether the change results in an admitted patient

waiting longer for an inpatient bed. We estimate Equation (1) with logged ED sojourn time and logged ED boarding time, respectively, as the dependent variable.

4. Results

4.1. Descriptive Statistics

Table 2(a) presents means and standard deviations for all continuous variables included in the empirical models, stratified by location (main ED or RCA) and time period (pre-PAS or post-PAS). Table 2(b) presents the correlations between all continuous variables included in the empirical models. Table 2(c) presents percentages for all categorical or binary variables in the empirical models stratified by location and time period. As shown in Table 2(a), the average LOS for a patient seen in the main ED is approximately three and a half hours, and it is about 50 minutes for a patient seen in the RCA. There are, on average, three or four physicians staffing the main ED during a given eight-hour period (i.e., morning shift, afternoon shift, overnight shift), and one physician staffing the RCA. None of the correlations between variables in the same regression model have levels close to or higher than 0.80, minimizing concerns about multicollinearity (see Table 2(b)). We also check for multicollinearity by calculating variance inflation factors (VIF). The largest VIF is 5.45 and the mean VIF is 2.52 (not shown), both of which fall well below the conventional threshold of 10, providing additional evidence that multicollinearity is not a concern (Wooldridge 2012). As Table 2(c) shows, nearly 75% of main ED patients are of ESI level 3, with the remainder being predominantly split between ESI levels 2 and 4. About 65% of main ED patients had a lab ordered compared to less than 9% of RCA patients.

As expected, patients' average LOS differs significantly by their acuteness. Although, for brevity, we do not display the numbers in a table, we find that for patients of ESI levels 2 to 5, the relationship between LOS and ESI level is a generally monotonically increasing one, with patients of a higher acuteness

Table 2(a) Summary Statistics of Continuous Variables Included in Models

	Main ED				RCA			
	Pre-PAS		Post-PAS		Pre-PAS		Post-PAS	
Variable	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1. *ED length of stay* (minutes)	212.7	210.7	210.3	227.3	46.6	65.1	46.8	61.3
2. *ED wait time* (minutes)	43.9	42.9	33.6	30.4	54.8	43.6	46.0	33.4
3. *Discharge rate* (patients/hour)	1.8	0.9	1.8	0.8	3.3	1.3	3.4	1.4
4. *Age* (years)	43.3	24.3	42.5	24.6	28.4	20.6	26.0	20.2
5. *MDs on shift*	3.4	0.9	3.7	1.0	1.0	0.2	1.0	0.3
6. *Current waiting count*	1.9	1.1	1.7	0.9	3.9	2.6	3.5	2.3
7. *Current patient count*	5.3	2.7	5.3	2.7	6.2	3.4	5.9	3.1
8. *Shift number*	115.1	72.0	335.5	151.5	135.7	78.4	373.8	126.8
9. *Shift duration* (hours)	9.7	1.5	9.2	1.3	10.2	1.2	10.0	1.0
10. *ED sojourn time* (minutes)	256.7	210.4	243.9	226.2	101.4	78.2	92.7	69.0
11. *ED boarding time* (minutes)	329.0	418.7	165.3	252.0	256.3	390.0	122.5	249.4

Notes. $N = 217{,}213$. Excludes all observations from August 2008 to account for an acclimation period.

Table 2(b) Correlations of Continuous Variables Included in Models

Variable	1	2	3	4	5	6	7	8	9	10	11
1. *ED length of stay* (minutes)	1										
2. *ED wait time* (minutes)	−0.13*	1									
3. *Discharge rate* (patients/hour)	−0.22*	0.22*	1								
4. *Age* (years)	0.30*	−0.12*	−0.16*	1							
5. *MDs on shift*	−0.05*	0.07*	0.13*	−0.02*	1						
6. *Current waiting count*	−0.19*	0.50*	0.52*	−0.18*	0.13*	1					
7. *Current patient count*	−0.04*	0.33*	0.45*	−0.06*	0.08*	0.61*	1				
8. *Shift number*	−0.04*	−0.07*	0.08*	−0.05*	0.21*	0.01*	0.004*	1			
9. *Shift duration* (hours)	−0.11*	0.11*	0.19*	−0.05*	0.12*	0.16*	0.10*	−0.04*	1		
10. *ED sojourn time* (minutes)	0.98*	0.05*	−0.15*	0.28*	−0.04*	−0.10*	0.02*	−0.06*	−0.09*	1	
11. *ED boarding time* (minutes)	0.40*	0.07*	−0.02*	0.07*	−0.02*	0.06*	0.01	−0.17*	−0.003*	0.41*	1

Notes. $N = 217{,}213$. Excludes all observations from August 2008 to account for an acclimation period.
*$p < 0.05$.

Table 2(c) Percent of Sample by Categorical and Binary Variables Included in Models

	Main ED		RCA	
Variable	Pre-PAS	Post-PAS	Pre-PAS	Post-PAS
ESI level 2	7.88	14.05	—	—
ESI level 3	74.10	73.70	—	—
ESI level 4	17.68	11.85	96.23	96.53
ESI level 5	0.34	0.40	3.77	3.47
ESI level 1 *patient present*	8.78	9.16	9.69	9.92
Trauma patient present	7.36	27.68	7.62	29.25
Morning shift	34.21	35.55	40.58	37.73
Afternoon shift	44.59	43.63	53.87	55.99
Overnight shift	21.20	20.82	5.55	6.28
2007	57.92	—	56.23	—
2008[a]	42.08	14.39	43.77	16.16
2009[a]	—	52.47	—	54.80
2010	—	33.15	—	29.04
January	5.78	8.73	6.35	8.59
February	6.18	8.45	6.63	8.18
March[a]	12.38	9.63	11.81	9.17
April[a]	11.74	9.18	11.18	8.70
May[a]	12.07	9.85	12.24	9.36
June[a]	11.63	8.93	11.25	8.45
July[a]	12.12	9.19	11.44	8.87
August[a]	5.88	4.50	5.96	4.72
September	5.66	8.10	5.72	8.68
October	5.55	8.01	5.77	9.17
November	5.46	7.72	5.75	8.34
December	5.55	7.69	5.89	7.76
Sunday	15.15	14.75	15.01	15.09
Monday	14.89	15.19	15.00	15.22
Tuesday	14.08	14.11	14.50	14.07
Wednesday	13.93	13.40	13.72	13.58
Thursday	13.84	13.86	13.89	13.40
Friday	13.84	13.95	13.37	13.33
Saturday	14.27	14.73	14.51	15.31
Lab ordered	64.12	66.91	8.83	8.05
x-ray ordered	38.44	39.46	27.37	26.92
Admitted to hospital	14.11	12.42	0.38	0.30
Revisit within 72 hours	4.99	5.04	2.89	2.77

Notes. $N = 217{,}213$. Excludes all observations from August 2008 to account for an acclimation period.

[a]Because the study period begins on March 1, 2007 and ends on July 31, 2010, it is not surprising that a larger percentage of patients in our data set presented to the ED in the months between March and July (inclusive) and in the years of 2008 and 2009, respectively. Because all observations from August 2008 have been excluded, it is also not surprising that this percentage is smaller for the month of August. When these summary statistics are produced with the inclusion of all observations from January 1, 2007, to December 31, 2010, we obtain an approximately uniform distribution of patients across all months of the year.

having a longer LOS. We account for the nonlinearity of this relationship by adjusting for patient acuteness using a dummy variable for each ESI level.

4.2. Patient Assignment System Implementation in the Main ED

Both the qualitative and quantitative data suggest that PAS was implemented as described, though not without challenges. An ED physician remarked on one of the key challenges during implementation: "[PAS] was the hardest thing we have ever done. When we first started with the PAS system, it was a rocky road because sometimes there were patients in the waiting room when there was an open bed." This comment, in combination with the first author's observations of the ED workflow, suggests that physicians largely abided by PAS and the RR routing policy. In our EMR data, we find further support for the general adherence to the RR routing policy. In particular, patient demographics across physicians are well balanced and there is little variation in the average acuteness of patients assigned to each physician, suggesting it is not the case that certain physicians are being assigned particular types of patients. Furthermore, on average, there are only one or two ESI level 2 patients seen by a physician on a given main ED shift (mean = 1.4, s.d. = 0.5), suggesting the workload across physicians remains relatively balanced, thus allowing physicians to feasibly adhere to the RR routing policy.

However, there are rare situations when the RR routing policy is violated. Although the internal triage nurse cannot bypass the patient assignment generated by the computer system, physicians working in the main ED can bypass the RR assignment determined by the computer when another physician has an exceptionally time-consuming workload of ESI level 1 patients. One physician stated the following: "The expectation is that each physician sees the patients assigned to him or her. Ninety-nine percent of the time, this happens . . . [but] we help each other if someone gets slammed with a critical [ESI level 1] patient . . . I remember one case last year where a physician got three critical patients in a row. That is extremely rare. He did not ask anyone, but two of his colleagues came and took two of the three patients [onto their panels]." This corroborates our understanding of the RR routing policy, in which other physicians can voluntarily take on the next patient assigned to a physician caring for an ESI level 1 patient.

4.3. Base Results

4.3.1. ED LOS. We estimate Equation (1) to assess the impact of having a pooled queuing system (versus a dedicated queuing system) on patients' average LOS in the main ED. Column (1) of Table 3 presents a fixed-effects model that captures the effect of moving from a pooled to a dedicated queuing system. We find that the difference in patients' average LOS between the main ED and the RCA is greater prior to PAS implementation. Once the main ED adopts a dedicated queuing system, this difference in patients' average LOS is reduced. This difference in differences is captured by the coefficient on the interaction term, $PAS \times main$ ($\alpha_2 = -0.17$, $p < 0.001$), and indicates that

Table 3 Fixed-Effects Models at Patient Level

Variables	(1) Logged ED length of stay	(2) Logged ED wait time
Main ED	0.642***	0.377***
	(0.0307)	(0.0330)
PAS × *Main ED*	−0.174***	−0.0854**
	(0.0211)	(0.0265)
ESI level 3	−0.401***	0.415***
	(0.0159)	(0.0129)
ESI level 4	−1.211***	0.698***
	(0.0203)	(0.0241)
ESI level 5	−1.578***	0.617***
	(0.0252)	(0.0291)
Age	0.00773***	−0.00260***
	(0.000233)	(0.000188)
MDs on shift	−0.00559	0.0148
	(0.00302)	(0.00855)
Current waiting count	0.00184	0.189***
	(0.00171)	(0.00561)
Current patient count	0.000909	0.0201***
	(0.00167)	(0.00476)
Shift number	−0.000484*	−2.51e−05
	(0.000236)	(0.000359)
ESI level 1 *patient present*	0.0169**	0.0604***
	(0.00528)	(0.00806)
Trauma patient present	0.00844	0.0650***
	(0.00505)	(0.00562)
Afternoon shift	−0.0605***	0.0726***
	(0.00711)	(0.0161)
Overnight shift	−0.0731***	−0.157***
	(0.0131)	(0.0283)
Constant	4.546***	2.298***
	(0.0538)	(0.0611)
Observations	217,161	217,213
Number of ED physicians	40	40
Adjusted R^2	0.519	0.298

Notes. All regressions are estimated at the patient level and include day of week controls, month-year fixed effects, and physician fixed effects. Standard errors (in parentheses) are heteroskedasticity robust and clustered by physician.

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

the transition from a pooled queuing system to a dedicated queuing system is associated with a highly significant reduction in the difference between the average LOS in the main ED and the RCA. This 17% decrease in the difference in average LOS in the main ED and the RCA after the implementation of PAS corresponds to a 39-minute decrease in LOS in the main ED relative to the RCA for an average patient of ESI level 3 seen by an average physician in the main ED. In other words, the average patient's LOS in the main ED when compared with that in the RCA is significantly longer in the pooled queuing system than in the dedicated queuing system. This result offers strong support for Hypothesis 1, which predicts that, in our setting, pooled queuing systems are associated with a longer average LOS compared to dedicated queuing systems.

This finding is consistent with strategic changes in physicians' behaviors to improve the management of their overall workflow. After PAS implementation, physicians change their practice behaviors because (a) they are aware of their full set of assigned patients, even those still in the waiting room, and (b) they have ownership over a designated bank of beds and nurses. In addition, when one of their designated beds becomes available because of a patient discharge, physicians post-PAS are responsible for ensuring that their next patient from the waiting room is placed in that bed as quickly as possible. Specifically, according to interviews with physicians and observations of their practice patterns, physicians change their practice behaviors by (a) proactively "pulling" for lab results, x-ray results, and specialty consult results rather than waiting for this information to be "pushed"; (b) jointly managing their own workflow with that of the nurses with whom they are paired to better coordinate various tasks; (c) initiating the discharge process sooner for patients who are ready for discharge; and (d) making sure that patients are brought in from the waiting room as soon as one of their main ED beds becomes available rather than waiting for the internal triage nurse to place the next patient in an open bed. Collectively, these proactive actions lead to a shorter average LOS for patients in the main ED and result in a decrease in the difference in average LOS between the main ED and the RCA.

To confirm that the implementation of PAS only affected the main ED and not the RCA, a necessary condition for using the difference-in-differences framework, we conduct two analyses. First, using a pre-post analysis that is limited to the RCA, we examine whether there is a discontinuous jump in LOS in the RCA when PAS is implemented. We find no evidence of a significant increase or decrease in LOS in the RCA after PAS implementation ($\alpha_2 = 0.01$, $p \approx 0.84$). Second, we check for a change in the slope of LOS trends in the RCA before and after PAS implementation. A Wald test on the equality of coefficients also suggests no change in the trend of LOS in the RCA after PAS implementation ($p \approx 0.71$). Both of these findings indicate that the effects of PAS implementation were limited to the main ED and did not affect the RCA, thereby validating the use of the difference-in-differences model.

4.3.2. ED Wait Time. We estimate Equation (2) to examine the impact of having a pooled queuing system on patients' average wait time in the main ED. The results are summarized in column (2) of Table 3. We find that the difference in patients' average wait time between the main ED and RCA decreases after PAS implementation ($\beta_2 = -0.09$, $p < 0.01$). This 9%

decrease corresponds to a four-minute decrease in wait time in the main ED relative to the RCA for an average patient of ESI level 3 seen by an average physician in the main ED. In other words, the average patient's wait time in the main ED when compared with that in the RCA is significantly longer in the pooled system than in the dedicated system. This offers strong support for Hypothesis 2A, which predicts that, in our setting, dedicated queuing systems are associated with a shorter average wait time compared to pooled queuing systems. We do not find support for Hypothesis 2B, which relies on traditional queuing theory to predict that a pooled queue yields a shorter average wait time than do dedicated queues. In the dedicated system, the shorter wait times may be attained because, instead of waiting for the internal triage nurse to initiate placing the next patient in an open bed, physicians operating under PAS are able to initiate placement of the next patient from their queue into their newly available bed. Our findings are also consistent with the expectation of an indirect queuing effect, where patients experience shorter wait times because the patients who are receiving care have a shorter average LOS, which in turn makes beds in the main ED available sooner.

4.3.3. Discharge Rate. To better understand how dedicated queuing systems impact patients' average LOS, we estimate Equation (3). We examine whether, and at what point during a physician's shift, the implementation of PAS affects the discharge rate of patients in the main ED.

Columns (1)–(4) of Table 4 present fixed-effects models estimated at the physician-shift two-hour period level for each of the following four time periods: the first, second, penultimate, and final two-hour periods of a physician's shift. We find that in the second, penultimate, and final two-hour periods of a shift, the discharge rate in the main ED exhibits a significant increase after PAS implementation. Specifically, after PAS implementation, the discharge rate is 1.05 times greater ($\varphi_1 = 1.05$, $p < 0.05$) in the second two hours, 1.07 times greater ($\varphi_1 = 1.07$, $p < 0.001$) in the penultimate two hours, and 1.05 times greater ($\varphi_1 = 1.05$, $p < 0.01$) in the final two hours of a physician's main ED shift. We also find that this increase in discharge rate does not manifest in the first two hours of a physician's main ED shift ($\varphi_1 = 1.04$, $p \approx 0.12$). Based on observations in the ED and the fact that the average LOS of a patient seen in the main ED is 211 minutes (i.e., approximately three and a half hours), the lack of significant difference in discharge rates in the first two hours of a shift may be due to the fact that the baseline amount of time necessary for patient care in the main ED is greater than two hours and, therefore, it is difficult for physicians to have a faster discharge rate during the first two hours of a shift.

Our findings are thus consistent with Hypothesis 3, which predicts that physicians in a dedicated queuing system exhibit a higher discharge rate that is sustained throughout the entire shift, which indicates that physicians are engaging in strategic behaviors over the entire course of the shift. This may be attributable to their greater ownership for patient flow and the resources needed to manage patient flow that comes with working in the ED's dedicated queuing system.

4.4. Consideration of Alternate Explanations and Unintended Consequences

Though our finding of a reduction in the difference between main ED and RCA patients' average LOS in a dedicated queuing system versus a pooled queuing system is consistent with an increase in physicians' strategic behavior to more efficiently manage patient flow, we consider alternate explanations that could also be consistent with our finding. We also explore the possibility of unintended consequences arising when implementing a dedicated queuing system.

4.4.1. Testing for Changes in the Provision of Care. First, one possibility is that physicians stint on care after PAS implementation because of the increased pressure to care for all patients in their dedicated queues. If fewer services are provided to patients, they may stay in the ED for a shorter amount of time. For example, if a patient who would have otherwise received an x-ray does not, she would likely stay in the ED for a shorter duration because she would not need to wait for the x-ray machine to become available, have the x-ray taken, and wait for the radiologist to read the films. If physicians are stinting on care post-PAS, we would be mistaken to assume that the reduced LOS stems from an increase in physicians' strategic behaviors to more efficiently manage patient flow.

We do not find strong evidence of stinting on care after the transition to a dedicated queuing system in the main ED. In column (1) of Table 5, we examine the change in a patient's likelihood of having a lab test ordered. We find that the coefficient for $PAS \times main$ is not statistically significant ($\alpha_2 = -0.08$, $p \approx 0.07$), suggesting that the difference in the likelihood of having a lab test ordered for a patient in the main ED and the RCA does not change significantly after PAS. Similarly, in column (2) of Table 5, we do not find a statistically significant change in a patient's likelihood of having an x-ray ordered ($\alpha_2 = -0.03$, $p \approx 0.50$). This suggests that there is no meaningful change in the difference in a patient's likelihood of receiving an x-ray between the main ED and the RCA before and after the implementation of PAS. In combination,

Table 4 Fixed-Effects Models at Physician-Shift Levels

Variables	(1) Discharge rate in first two hours of shift	(2) Discharge rate in second two hours of shift	(3) Discharge rate in penultimate two hours of shift	(4) Discharge rate in last two hours of shift	(5) Shift duration
Main ED	—	—	—	—	1.060***
					(0.166)
PAS	1.042	1.053*	1.069***	1.051**	—
	(0.0275)	(0.0245)	(0.0190)	(0.0204)	
PAS × *Main ED*	—	—	—	—	−0.0904
					(0.0855)
Percent of ESI level 3 *patients*	1.001	1.002***	1.002***	1.001*	0.0137***
	(0.000641)	(0.000430)	(0.000440)	(0.000404)	(0.00162)
Percent of ESI level 4 *patients*	1.007***	1.008***	1.005***	1.004***	0.0325***
	(0.000761)	(0.000535)	(0.000668)	(0.000553)	(0.00233)
Percent of ESI level 5 *patients*	1.012***	1.007*	1.001	1.006*	0.0311***
	(0.00368)	(0.00323)	(0.00299)	(0.00240)	(0.00537)
Average age of patients	1.000	1.000	1.001***	1.000	−0.0208***
	(0.000269)	(0.000267)	(0.000295)	(0.000386)	(0.00309)
MDs on shift	0.982*	0.981*	0.976**	0.968***	−0.331***
	(0.00736)	(0.00884)	(0.00730)	(0.00662)	(0.0239)
Average waiting count	1.015	1.014	0.999	1.001	−0.530***
	(0.0184)	(0.00961)	(0.00925)	(0.00694)	(0.0364)
Average patient count	1.094***	1.123***	1.111***	1.104***	0.661***
	(0.0172)	(0.00861)	(0.00510)	(0.00432)	(0.0260)
Shift number	1.000	1.000	1.000	1.000	7.06e–05
	(0.000231)	(0.000226)	(0.000184)	(0.000142)	(0.000805)
Percent of time ESI level 1 *patient present*	0.996	0.995	0.980	1.022	0.0495
	(0.0279)	(0.0186)	(0.0172)	(0.0161)	(0.0541)
Percent of time trauma patient present	0.998	0.965*	1.000	1.016	−0.0234
	(0.0185)	(0.0157)	(0.00994)	(0.0132)	(0.0479)
Afternoon shift	0.936***	1.064**	1.154***	1.156***	−0.134*
	(0.0133)	(0.0208)	(0.0199)	(0.0152)	(0.0541)
Overnight shift	0.806***	1.031	1.094***	0.980	−1.834***
	(0.0203)	(0.0382)	(0.0225)	(0.0235)	(0.130)
Constant	—	—	—	—	6.917***
					(0.301)
Observations	3,922	8,594	10,675	10,905	14,153
Number of ED physicians	38	39	38	40	40
Adjusted R^2	—	—	—	—	0.329

Notes. Columns (1)–(4) are conditional fixed-effects Poisson models estimated at the physician-shift two-hour period level with linear time trends by month, day of week controls, physician fixed effects, and heteroskedasticity robust standard errors. Discharge rate reflects the number of patients discharged per hour by a given physician in a given two-hour period of the shift, and coefficients have been exponentiated to show incident rate ratios. Column (5) is a fixed-effects linear regression model estimated at the physician-shift level with day of week controls, month-year fixed effects, physician fixed effects, and heteroskedasticity robust standard errors clustered by physician. Shift duration is expressed in hours.

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

these results suggest that physicians are not systematically stinting on care in the main ED as compared to the RCA as a result of PAS implementation.

4.4.2. Testing for Changes in the Likelihood of a Patient's Admission to Hospital. A second possibility is that ED physicians may be reducing patients' average LOS in the ED by passing them off to other hospital departments earlier. If an ED physician decides to have a patient admitted to the inpatient unit for further evaluation, rather than taking the time to conduct further evaluation while the patient is still in the ED, the patient's LOS in the ED may appear to be shorter than it would be otherwise.

We do not find evidence of main ED patients exhibiting a higher likelihood of admission to the hospital, relative to RCA patients, after PAS. As shown in column (3) of Table 5, we find that the difference in a patient's likelihood of being admitted to the hospital when in the main ED versus the RCA does not change significantly after PAS implementation ($\alpha_2 = -0.19$, $p \approx 0.16$).

4.4.3. Testing for Changes in the Quality of Care. Next, we consider two potential unintended consequences of this transition from a pooled to a dedicated queuing system in the main ED. We assess whether patients are more likely to return to the ED within

Table 5 Logistic Regression Models at Patient Level for Alternate Explanations and Unintended Consequences

Variables	(1) Lab ordered	(2) x-ray ordered	(3) Admitted to hospital	(4) Revisit within 72 hours	(5) Died in ED
Main ED	1.451***	−0.103*	1.673***	0.167***	—
	(0.120)	(0.0503)	(0.102)	(0.0446)	
PAS	—	—	—	—	−0.669*
					(0.301)
PAS × *Main ED*	−0.0847	−0.0260	−0.188	0.00770	—
	(0.0468)	(0.0382)	(0.132)	(0.0506)	
ESI level 2	—	—	—	—	−5.270***
					(0.420)
ESI level 3	−0.693***	−0.508***	−1.007***	0.0381	−7.457***
	(0.0319)	(0.0334)	(0.0322)	(0.0375)	(0.538)
ESI level 4	−2.550***	−0.799***	−2.929***	−0.374***	−8.820***
	(0.0430)	(0.0476)	(0.0804)	(0.0633)	(1.008)
ESI level 5	−3.275***	−2.348***	−5.300***	−0.577***	—
	(0.0732)	(0.117)	(0.994)	(0.155)	
Age	0.0176***	0.0221***	0.0389***	0.00125*	0.0284***
	(0.000652)	(0.000937)	(0.000692)	(0.000515)	(0.00435)
MDs on shift	−0.0146	−0.0205	−0.00650	−0.0106	−0.0915
	(0.0180)	(0.0113)	(0.0127)	(0.0136)	(0.0974)
Current waiting count	0.0131	0.00726	−0.0187	−0.0232**	−0.0480
	(0.00723)	(0.00572)	(0.0111)	(0.00803)	(0.0750)
Current patient count	−0.0165***	0.00217	−9.74e−05	−0.00847	−0.00174
	(0.00462)	(0.00381)	(0.00435)	(0.00436)	(0.0307)
Shift number	−0.000511	−0.000515*	−0.000314	8.50e−05	−7.19e−05
	(0.000303)	(0.000260)	(0.000265)	(0.000122)	(0.000688)
ESI level 1 *patient present*	0.0316	−0.0101	−0.0432	0.00331	−0.0577
	(0.0276)	(0.0163)	(0.0288)	(0.0467)	(0.439)
Trauma patient present	0.0117	0.0241	−0.00760	0.0513	−0.204
	(0.0183)	(0.0148)	(0.0286)	(0.0369)	(0.174)
Afternoon shift	−0.0999	0.0286	0.0460	−0.00747	0.226
	(0.0552)	(0.0274)	(0.0338)	(0.0327)	(0.162)
Overnight shift	−0.175***	−0.0177	0.000590	0.0713	0.459
	(0.0532)	(0.0330)	(0.0472)	(0.0559)	(0.282)
Constant	−0.558***	−0.767***	−4.317***	−3.035***	−1.687*
	(0.160)	(0.0771)	(0.122)	(0.131)	(0.729)
Observations	193,807	193,807	193,807	193,807	132,952
Pseudo R^2	0.331	0.0679	0.257	0.0110	0.564

Notes. All regressions are logistic regression models estimated at the patient level. Columns (1)–(4) include day of week controls, month-year fixed effects, and physician fixed effects. Column (5) includes linear time trends by month, day of week controls, and physician fixed effects. Column (5) also includes previously excluded observations—specifically patients of ESI level 1, patients who died in the ED, and trauma patients. Standard errors (in parentheses) are heteroskedasticity robust and clustered by physician.

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

72 hours of being seen, which could be an unintended consequence of physicians providing lower quality or insufficient care in order to decrease patient LOS. Similarly, if physicians are providing lower quality care such that more patients are dying in the ED, this truncating effect on LOS may result in a decrease in the average patient's LOS in the ED.

We do not find evidence of a lower quality of care as measured by revisits to the ED within 72 hours. As is summarized in column (4) of Table 5, we find no statistically significant changes after PAS implementation in the difference in the likelihood of returning to the ED within 72 hours of an initial visit ($\alpha_2 = 0.01$, $p \approx 0.88$). Even when using a more inclusive cutoff of seven days (results not shown), we find no statistically significant changes in the difference in the likelihood of revisit ($\alpha_2 = 0.07$, $p \approx 0.11$).

In addition, we do not find evidence of a lower quality of care as measured by mortality in the ED. These results are presented in column (5) of Table 5. Because of the lack of variation in the dependent variable among patients of ESI level 5 and patients seen in the RCA, these two categories of patients are omitted from the analysis. In the resulting analysis, comparing

patient mortality in the main ED before and after the implementation of PAS, we find that the likelihood of dying in the ED decreased after the transition to a dedicated queuing system ($\alpha_2 = -0.67$, $p < 0.05$). This suggests that the quality of care, as measured by patient mortality in the ED, *improved* after PAS was implemented, thereby reducing concerns that the assignment of patients in the waiting room to a specific physician might adversely affect patients.

4.4.4. Testing for Potential Impact on the Duration of a Physician's Shift. Finally, we consider the potential impact of PAS on the duration of a physician's shift. As summarized in column (5) of Table 4, we find no statistically significant change in the difference between the duration of a shift in the main ED and the RCA before and after PAS implementation ($\alpha_2 = -0.09$, $p \approx 0.30$). This suggests that physicians are not working longer hours in the main ED as a result of the intervention.

4.5. Specification Tests

To examine the robustness of our main findings about LOS, we test a variety of other specifications in addition to the reported models. Because of space constraints, these results are not reported in the tables.

First, we use a limited model specification that includes only patient ESI levels as control variables. We retain patient ESI levels because the average acuteness of patients arriving in the main ED increased over time, whereas that of patients arriving in the RCA decreased over time. We find that the base result remains very robust to this limited model specification ($\alpha_2 = -0.16$, $p < 0.001$), with the magnitude of the effect decreasing only slightly from 17% to 16%.

We then repeat our estimation of Equation (1) using nonlogged LOS and bootstrapped standard errors. With this alternate model specification, we find that PAS implementation is associated with a 23-minute reduction in the difference in LOS between the main ED and the RCA ($\alpha_2 = -22.73$, $p < 0.001$). Even when not using a log-level specification to account for the heavily skewed nature of the dependent variable, we obtain results that are robust to our base findings.

Although our interviews with ED staff suggest that there were no other interventions besides PAS that were applied to only the main ED or only the RCA during the study period (March 1, 2007 to July 31, 2010), we apply our analyses to shorter time frames around PAS implementation to nullify the possibility of other effects. When we limit the time frame to three months, seven months, 12 months, 15 months, and 18 months before and after the intervention, we find that our base results remain robust to these shorter time frames ($\alpha_2 < -0.10$, $p < 0.001$).

Next, we repeat our analyses using logged ED sojourn time to test for the impact of PAS on a more holistic measure of patient experience. We find that PAS is associated with a 10% decrease in the difference between main ED and RCA sojourn times before and after PAS implementation ($\alpha_2 = -0.10$, $p < 0.001$). This suggests that when taking both wait time and LOS into account, PAS is associated with a reduction in the average time that patients spend in the ED. In addition, we examine the impact of PAS on logged ED boarding time, which is the amount of time that patients being admitted to the hospital spend waiting for an inpatient bed. We find no statistically significant change in the difference in ED boarding times for patients in the main ED and the RCA before and after PAS implementation ($\alpha_2 = -0.25$, $p \approx 0.09$). This is consistent with our expectation because ED boarding time is primarily determined by the inpatient unit's capacity to admit a new patient rather than ED physicians' productivity levels.

Next, we limit our sample to those patients seen in the main ED and conduct a pre-post analysis, comparing the average LOS of patients before and after PAS. We find that our main findings are robust to this alternate specification that does not use a difference-in-differences approach, where PAS is associated with a 5% decrease in LOS in the main ED ($\alpha_2 = -0.05$, $p < 0.01$).

In addition, our results do not appear to be driven by differences in patient care delivered in the two areas of the ED. To examine this, we assess whether the transition from a pooled system to a dedicated system differentially affects LOS depending on the location of a patient's ED care. To conduct this analysis, we use the same empirical model as Equation (1), but limit the sample to patients of ESI levels 4 and 5, and with each independent variable of interest interacted with ESI level 5. We limit the sample to these patients because they constitute the group of patients who are potentially seen in both areas of the ED (because all ESI levels 4 and 5 patients are seen in the main ED after 11 p.m.). This analysis suggests that there are no differential effects by the location of a patient's ED care ($p \approx 0.32$).

Furthermore, we examine whether the base results are sensitive to heterogeneity in patient acuteness. In other words, we examine whether the transition from having a pooled queuing system to a dedicated queuing system has a greater impact on patients with a higher ESI level as opposed to those with a lower ESI level. Using a similar approach as above, we explore this possibility by limiting the sample to patients of ESI levels 2 and 3, and interacting each independent variable of interest with ESI level 3. For this analysis, we limit the sample to patients of these two ESI levels because they exhibit two different groups with relatively longer LOS (for ESI level 2, mean = 332 minutes, s.d. = 330 minutes) and shorter LOS (for

ESI level 3, mean = 212 minutes, s.d. = 202 minutes). This analysis suggests that patients of higher acuteness (ESI level 2) are likely to experience a greater decrease in LOS after the implementation of PAS compared with patients of a relatively lower acuteness (ESI level 3) ($\alpha = 0.42$, $p < 0.01$). Although it is beyond the scope of this paper to examine why this heterogeneity arises, we speculate that it may be due to the prioritization of higher acuteness patients (ESI level 2) within each physician's dedicated queue.

We also repeat our analyses using several different exclusion criteria in constructing our sample and find that our results are robust in all of the following analyses. First, we include all observations that had previously been excluded as outliers (i.e., patients with a LOS greater than 48 hours). Then, to test our hypotheses on an even more homogeneous set of patients and ensure that our findings are not driven by outliers, we exclude observations with a LOS greater than one day (24 hours) and the average duration of one shift (9.4 hours), respectively. Next, we test our hypotheses on a sample that includes ESI level 1 patients, which were previously excluded. Finally, we test our hypotheses on a sample that excludes patients arriving by ambulance and patients presenting with a psychological condition, respectively, both of whom were previously included. All coefficients of interest and their corresponding significance levels remain robust to these alternate specifications ($\alpha_2 < -0.13$, $p < 0.001$).

Finally, we use hierarchical linear models, which specify random effects rather than fixed effects at the physician level. We conduct this analysis to test each of our hypotheses with greater efficiency gains. We use three levels for our multilevel analyses: patient, physician-shift, and physician. The effect of transitioning from having a pooled queuing system to a dedicated queuing system remains robust to this model specification ($\alpha_2 = -0.18$, $p < 0.001$).

5. Discussion and Conclusions

Using three and a half years of data from a hospital's ED, we find that patients experience shorter LOS when physicians work in a dedicated queuing system with a fairness constraint as opposed to a pooled queuing system with the same fairness constraint. Although we are unable to precisely test the mechanism for the shorter LOS in the dedicated system, we believe that the improved performance stems from strategic physician behaviors triggered by physicians' greater ownership over patient flow and the resources needed to smooth flow through the ED. This suggests that the flow management benefits associated with a dedicated queuing system with a fairness constraint may outweigh the variability-buffering benefits of a pooled queuing system. We consider, but find no empirical support for, alternate explanations for this reduction in LOS, such as changes in the provision of care or lower quality care.

We find evidence that physicians' strategic behaviors persist throughout the entire shift. In particular, examination of physicians' discharge rates in two-hour periods over the course of the shift shows that physicians exhibit a higher discharge rate when working in a dedicated queuing system as opposed to a pooled queuing system soon after beginning the shift. This increase in discharge rates is sustained throughout the remainder of the shift. In describing how the implementation of PAS increases physicians' ability to manage patient flow, one physician said,

> Before PAS, the physician had no control or responsibility over getting the next patient into an empty bed. I often had idle time and had more than enough time to see more patients; I just couldn't get them to me from the waiting room. I wasn't in control so I didn't do much to get patient turnover to happen faster. Now, with PAS, I am responsible for getting my patients from the waiting room into my beds. I do this by making sure that tasks are being done so that I can discharge my current patients.... It changed the whole responsibility for patient flow from [the] one [internal triage] nurse onto me to manage my patients.

To quantify the impact of our findings, we calculate effect sizes. We find that moving from a pooled queuing system to a dedicated queuing system is associated with a 17% decrease in the difference in LOS between the main ED and RCA. For an average patient of ESI level 3 seen in the main ED by an average physician, this corresponds to a 39-minute decrease in LOS in the main ED relative to the RCA. This is a particularly meaningful difference in the context of a hospital's emergency room. With approximately 200 patients in the ED every day, this is roughly equivalent to an additional 130 patient-hours per day that are saved with the dedicated queuing system. Once we take into account the large costs associated with emergency room care, it becomes clear that the time and cost implications are substantial. If these findings are generalizable to other EDs, this would have significant practical implications for EDs across the country faced with large increases in patient volume accompanied by constrained budgets.

Nevertheless, it is important to consider the potential limitations of dedicated queuing systems. In systems with less homogenous patient populations, a dedicated queuing system with fairness constraints might result in imbalanced workloads among different care providers.

5.1. Theoretical Contributions

This paper contributes to the operations management literature on queue pooling in several ways. Our paper

is one of a few to use empirical data to examine the effect of queue management systems on wait times and service times. We find that when servers have ownership over patient flow and key resources, dedicated queuing systems with a fairness constraint are associated with *shorter* wait times and service times than pooled systems with a fairness constraint. Our findings illustrate the importance of accounting for the interaction between human behavior and queuing system design when predicting performance (Boudreau et al. 2003, Jouini et al. 2008). When queuing theory does not account for strategic server behavior, it suggests that pooling queues should result in shorter wait times even when fair routing policies are used (Armony and Ward 2010). In our study, we find that wait times are longer for the pooled system. Thus, our paper provides empirical support for prior analytical models that predict that human behaviors can reduce the benefits of using a common pool (Best et al. 2015, Cachon and Zhang 2007, Gilbert and Weng 1998, Hopp et al. 2007, Jouini et al. 2008, Wang et al. 2010). We are also able to add quantitative, empirical evidence to the debate that the benefits that arise from lean manufacturing's practice of assigning a specific person to service a specific stream of work outperforms the flexibility benefits from a pooled system (Spear and Bowen 1999). Our paper demonstrates how employees' willingness and ability to manage flow create an advantage for dedicated systems over pooled systems.

We speculate that queue pooling results in longer LOS because, in the pooled system, physicians do not feel completely responsible for patient flow because the internal triage nurse is responsible for moving patients from the waiting room to available beds. This result is similar to, but distinct from, Chan's (2015) finding that ED physicians work slower when they are assigned patients by a triage nurse than when physicians—collectively as a group—assign patients to physicians. Chan asserts that this "foot-dragging" behavior occurs in the nurse-managed system because physicians delay discharges to overstate their true workload to the nurse in hopes of avoiding being assigned another patient. The findings in the Debo et al. (2008) study are also driven by servers' misleading behaviors. Another mechanism in the literature that explains why dedicated queuing systems have faster service times than pooled systems is that managers can better supervise the smaller teams of workers that result from splitting up a large pooled system into a set of dedicated systems and a healthy competition emerges among the different dedicated systems (Jouini et al. 2008).

In contrast, we propose a different underlying mechanism for the improvement in throughput times: better flow management arising from strategic physician behaviors. In our study, a computer-automated RR routing policy fairly assigns patients to physicians both before and after the intervention. Thus, unlike physicians in Chan's (2015) study, physicians in our study are not deliberately working slower to overstate their workloads. Furthermore, the fact that only a handful of physicians are working in this ED at any one time suggests that the Jouini et al. (2008) emphasis on the challenge of managing a large pool of employees is not what is driving our results. Also, physicians were not given any information about other physicians' average LOS, so competition is not the explanatory mechanism (Jouini et al. 2008). Instead, we propose that making a single physician—as opposed to a group of physicians—accountable for efficiently managing patient flow leads to a reduction in wait time and LOS through better flow management practices.

Our findings build on the Schultz et al. (1998) study of the motivational impact of low inventory levels on production line workers' speeds. Schultz et al. (1998) find that low inventory motivates slower workers to speed up, enough to cancel the productivity loss due to the blocking and starving that occurs in low inventory production lines. We examine a different lever to increase workers' motivation: the queue structure of incoming jobs. We find that, when physicians work in a dedicated queuing system, they are able to attain shorter average LOS and wait times for their patients by managing their workloads more efficiently. We suggest that this may be because the dedicated system affords physicians a higher level of ownership over patient flow. We find that the motivational benefits of the dedicated queuing system outweigh the inefficiencies introduced by unpooling the queue. Thus, our study furthers the Schultz et al. (1998) finding by proposing that queue structure is another job design factor that interacts with human behavior in ways that can reverse predicted relationships between work system design and performance.

5.2. Implications for Practice

Our study has important implications for workplace managers and healthcare policy makers. Our findings suggest that managers of work settings with strategic servers should design work systems to mitigate behaviors that benefit the employee to the detriment of customers or the organization. One possible mechanism is to give strategic servers greater ownership and responsibility for managing their workflow and to route work evenly across all servers regardless of differences in work pace, which removes the benefit of working slower than one's peers. EDs may benefit from implementing dedicated, fair queuing systems in which patients are assigned to physicians immediately following triage. To our knowledge, this is not currently in place at most EDs; most EDs employ a

pooled queuing system that assigns patients to physicians once placed in a bed. Thus, the potential for improvement is significant.

5.3. Limitations and Future Research

This study has limitations, and its results should be interpreted accordingly. First, we note the threat of omitted variable bias, common to many empirical models. Although it would have been helpful to include more patient characteristics in our model, such as patient diagnoses or medical comorbidities, these data were protected information and not available for use. However, this is not an important threat to validity because patients are randomly assigned to physicians rather than by physician choice. This is supported by the fact that the average ESI level of patients seen by each physician is less than one standard deviation away from the average ESI level of all patients seen in the ED (mean = 3.33, s.d. = 0.64).

Second, our study is of a single hospital's ED and its response to a single intervention. The fact that our data come from a single ED makes it impossible for us to use another ED as the control in our difference-in-differences analysis. Although we are confident that the RCA is a good control for our study, there would be advantages to using data from another ED with a similar patient population that did not implement the PAS system. We were unable to do this in our study because the PAS system was implemented in all EDs in the hospital's network. Though the generalizability of our findings is limited because we studied only one ED, we believe our findings have strong theoretical underpinnings. Nevertheless, future research could examine a larger sample of EDs to study a wider variety of routing policies and queue structures. Given that prior literature has found a variety of different mechanisms that may explain the shorter service times in dedicated systems, such research might enable greater clarity in which mechanisms are most powerful and under what conditions. In addition, these effects and suggested mechanisms could be studied in different empirical contexts for further theory development.

Third, our study raises the possibility that better flow management—arising from ownership over key resources—enables physicians in dedicated queuing systems to reduce their patients' average wait times and LOS. However, we are unable to precisely identify and test the mechanisms conclusively. Instead, we suggest these potential mechanisms based on interviews with physicians and observations of their practice patterns and leave it to future research to disentangle the mechanisms responsible for the reduced times.

Fourth, future research could consider how dedicated queuing systems affect patient and physician satisfaction, since changes in wait times and LOS may be associated with perceptions of fairness and the general satisfaction of both parties. These data are not available from the time period of our study, but have recently become more widely available.

Finally, implementing a dedicated queuing system is merely one way to try to attain the goal of shorter wait times and LOS in EDs. Future research should consider other mechanisms, such as financial incentives or interventions that leverage social pressure (Chan 2015). For example, do physicians increase their work rates when provided information about each other's average LOS? It may be possible to use a combination of interventions so that EDs can capture the benefits of pooling while simultaneously avoiding the slower service rates that seem to arise from queuing systems where responsibility for customers is shared across multiple servers.

5.4. Conclusions

Effectively using queue design to create both fairness and efficiency is an important opportunity for service organizations. Although results may differ across different settings, the mechanisms through which changes in LOS occur may help shed light on improvement opportunities in other contexts. Our findings are especially timely and could have significant implications for healthcare delivery as EDs across the country contemplate ways to handle the anticipated increases in ED patient volume as a result of the recent health reform legislation (Patient Protection and Affordable Care Act 2010, Pub. L. 111–148).

Acknowledgments

This research would not have been possible without the collaboration of Kaiser Permanente Northern California. In particular, the authors thank Mark B. Kauffman for his support and Brent E. Soon for his assistance in preparing the ED data. The authors thank Gérard P. Cachon, Laurens G. Debo, Wallace J. Hopp, Robert S. Huckman, Alexandra A. Killewald, Rajiv Kohli, Avishai Mandelbaum, Nirup Menon, Charles Noon, Tom Tan, Jan A. Van Mieghem; participants in the Longitudinal Data Analysis course at Harvard University; seminar participants at the 2013 INFORMS Healthcare Conference, the 2013 INFORMS MSOM Conference, the 73rd Annual Meeting of the Academy of Management, the 2013 INFORMS Annual Meeting, the Harvard Health Policy Research Seminar, and the 2014 AcademyHealth Annual Research Meeting; and the editor, associate editor, and three anonymous reviewers for their insightful comments. The authors also thank Simo Goshev, Tomoko Harigaya, Andrew Marder, and William B. Simpson for their advice regarding data analysis methods and Lydia Ypsse Kim for her expert research assistance. The authors acknowledge support for this research from the Division of Research and Faculty Development at Harvard Business School.

References

Abadie A (2005) Semiparametric difference-in-differences estimators. *Rev. Econom. Stud.* 72(1):1–19.

Anupindi R, Chopra S, Deshmukh SD, Van Mieghem JA, Zemel E (2005) *Managing Business Process Flows: Principles of Operations Management*, 2nd ed. (Prentice-Hall, Upper Saddle River, NJ).

Armony M, Ward AR (2010) Fair dynamic routing in large-scale heterogeneous-server systems. *Oper. Res.* 58(3):624–637.

Ata B, Van Mieghem JA (2008) The value of partial resource pooling: Should a service network be integrated or product-focused? *Management Sci.* 55(1):115–131.

Bassamboo A, Randhawa RS, Van Mieghem JA (2010) Optimal flexibility configurations in newsvendor networks: Going beyond chaining and pairing. *Management Sci.* 56(8):1285–1303.

Benjaafar S (1995) Performance bounds for the effectiveness of pooling in multi-processing systems. *Eur. J. Oper. Res.* 87(2):375–388.

Best TJ, Sandıkçı B, Eisenstein DD, Meltzer DO (2015) Managing hospital inpatient bed capacity through partitioning care into focused wings. *Manufacturing Service Oper. Management* 17(2):157–176.

Boudreau J, Hopp WJ, McClain JO, Thomas LJ (2003) On the interface between operations and human resources management. *Manufacturing Service Oper. Management* 5(3):179–202.

Cachon GP, Zhang F (2007) Obtaining fast service in a queueing system via performance-based allocation of demand. *Management Sci.* 53(3):408–420.

Chan DC (2015) Teamwork and moral hazard: Evidence from the emergency department. *J. Political Econom.* Forthcoming.

Debo LG, Toktay LB, Van Wassenhove LN (2008) Queuing for expert services. *Management Sci.* 54(8):1497–1512.

Deo S, Jain A, Pendem P (2014) Pacing work in the presence of goals and deadlines: Econometric analysis of an outpatient department. Working paper, Indian School of Business, Hyderabad, India.

Doroudi S, Gopalakrishnan R, Wierman A (2011) Dispatching to incentivize fast service in multi-server queues. *ACM SIGMETRICS Perform. Eval. Rev.* 39(3):43–45.

Duflo E (2001) Schooling and labor market consequences of school construction in Indonesia: Evidence from an unusual policy experiment. *Amer. Econom. Rev.* 91(4):795–813.

Eppen GD (1979) Note—Effects of centralization on expected costs in a multi-location newsboy problem. *Management Sci.* 25(5):498–501.

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* 5(2):79–141.

Gilbert SM, Weng ZK (1998) Incentive effects favor nonconsolidating queues in a service system: The principal-agent perspective. *Management Sci.* 44(12):1662–1669.

Green LV, Nguyen V (2001) Strategies for cutting hospital beds: The impact on patient service. *Health Services Res.* 36(2):421–442.

Hackman JR, Oldham GR (1976) Motivation through the design of work: Test of a theory. *Organ. Behav. Human Performance* 16(2):250–279.

Hasija S, Pinker E, Shumsky RA (2010) Work expands to fill the time available: Capacity estimation and staffing under Parkinson's law. *Manufacturing Service Oper. Management* 12(1):1–18.

Hopp WJ, Iravani SMR, Liu F (2009) Managing white-collar work: An operations-oriented survey. *Production Oper. Management* 18(1):1–32.

Hopp WJ, Iravani SMR, Yuen GY (2007) Operations systems with discretionary task completion. *Management Sci.* 53(1):61–77.

Hyytiä E, Aalto S (2013) Round-robin routing policy: Value functions and mean performance with job- and server-specific costs. *Proc. 7th Internat. Conf. Performance Evaluation Methodologies and Tools (ValueTools '13)* (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, Brussels), 69–78.

Jouini O, Dallery Y, Nait-Abdallah R (2008) Analysis of the impact of team-based organizations in call center management. *Management Sci.* 54(2):400–414.

Keith KD, Bocka JJ, Kobernick MS, Krome RL, Ross MA (1989) Emergency department revisits. *Ann. Emergency Medicine* 18(9):964–968.

Kleinrock L (1976) *Queueing Systems, Volume 2: Computer Applications* (John Wiley & Sons, New York).

Link S, Naveh E (2006) Standardization and discretion: Does the environmental standard ISO 14001 lead to performance benefits? *IEEE Trans. Engrg. Management* 53(4):508–519.

Loch C (1998) Operations management and reengineering. *Eur. Management J.* 16(3):306–317.

Mandelbaum A, Reiman MI (1998) On pooling in queueing networks. *Management Sci.* 44(7):971–981.

McCarthy ML, Ding R, Pines JM, Terwiesch C, Sattarian M, Hilton JA, Lee J, Zeger SL (2012) Provider variation in fast track treatment time. *Medical Care* 50(1):43–49.

Oliva R, Sterman JD (2001) Cutting corners and working overtime: Quality erosion in the service industry. *Management Sci.* 47(7):894–914.

Raz D, Avi-Itzhak B, Levy H (2006) Fairness considerations of scheduling in multi-server and multi-queue systems. *Proc. 1st Internat. Conf. Performance Evaluation Methodolgies and Tools (ValueTools '06)* (Association for Computing Machinery, New York), Article 39.

Rothkopf MH, Rech P (1987) Perspectives on queues: Combining queues is not always beneficial. *Oper. Res.* 35(6):906–909.

Schultz KL, Juran DC, Boudreau JW, McClain JO, Thomas LJ (1998) Modeling and worker motivation in JIT production systems. *Management Sci.* 44(12):1595–1607.

Spear S, Bowen HK (1999) Decoding the DNA of the Toyota production system. *Harvard Bus. Rev.* 77(5):96–106.

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

van Dijk NM, van der Sluis E (2008) To pool or not to pool in call centers. *Production Oper. Management* 17(3):296–305.

van Dijk NM, van der Sluis E (2009) Pooling is not the answer. *Eur. J. Oper. Res.* 197(1):415–421.

Wang X, Debo LG, Scheller-Wolf A, Smith SF (2010) Design and analysis of diagnostic service centers. *Management Sci.* 56(11):1873–1890.

Wooldridge JM (2010) *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. (MIT Press, Cambridge, MA).

Wooldridge JM (2012) *Introductory Econometrics: A Modern Approach*, 5th ed. (South-Western Cengage Learning, Mason, OH).