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

HARVARD | BUSINESS | SCHOOL

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

Hummy Song
Anita L. Tucker
Karen L. Murrell

Working Paper

14-050

December 17, 2013

The Diseconomies of Queue Pooling:

An Empirical Investigation of Emergency Department Length of Stay

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

Anita L. Tucker
Harvard Business School
Boston, MA 02163
atucker@hbs.edu

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

Revised: December 17, 2013

Acknowledgments

This research would not have been possible without the collaboration of the Kaiser Permanente South Sacramento Medical Center's Emergency Department (KP SSC ED). In particular, the authors are deeply appreciative of the support provided by Mark B. Kauffman, Director of Emergency Systems and Delivery Systems Optimization at KP SSC ED. The authors are also grateful to Brent E. Soon, Senior Financial Analyst at KP SSC, for providing us access to the dataset used in this paper. The authors thank Gerard 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, and the Harvard Health Policy Research Seminar; and the editor, associate editor, and three anonymous reviewers for their insightful comments. The authors also thank William B. Simpson, Simo Goshev, and Andrew Marder for their advice regarding data analysis methods and Lydia Ypsse Kim for her expert research assistance.

Abstract

We conduct an empirical investigation of the impact of two different queue management systems on throughput times. Using an Emergency Department's (ED) patient-level data (N = 231,081) from 2007 to 2010, we find that patients' lengths of stay (LOS) were longer when physicians were assigned patients under a pooled queuing system, compared to when each physician operated under a dedicated queuing system. The dedicated queuing system resulted in a 10 percent decrease in LOS—a 32-minute reduction in LOS for an average patient of medium severity in this ED. We propose that the dedicated queuing system yielded shorter throughput times because it provided physicians with greater ability and incentive to manage their patients' flow through the ED from arrival to discharge. Consistent with social loafing theory, our analysis shows that patients were treated and discharged at a faster rate in the dedicated queuing system than in the pooled queuing system. We conduct additional analyses to rule out alternate explanations, such as stinting on care and decreased quality of care. Our paper has implications for health care organizations and others seeking to reduce throughput time, resource utilization, and costs.

Key words: pooling, queue management, strategic servers, social loafing, empirical operations, health care

1. Introduction

Improving efficiency and customer experience are key objectives for managers of service organizations. Skillful application of operations management principles may prove helpful for achieving these goals. In this paper, we investigate queue management, a key operational decision. More specifically, we explore the impact on throughput time of using a pooled queuing system or a dedicated queuing system.

Prior work on queue management has demonstrated through analytical models that pooling separate streams of identical customers into a single queue served by a bank of identical servers leads to a reduction in waiting time and an increase in server utilization (Eppen 1979, Kleinrock 1976). This occurs because pooling reduces the negative impact of variability in arrival times and processing times. Pooling enables incoming work to be processed by any one server from a bank of servers, which decreases the odds that an incoming unit of work will have to wait for service. This situation compares to the one where the unit of work can only be processed by a single dedicated server.

Queuing theorists have advanced this stream of research by identifying conditions under which queue pooling may not 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). However, few empirical studies have examined the performance of pooled versus dedicated queuing systems. This is an important omission because, in practice, employees can often make adjustments to how they manage their work system to achieve a goal, such as increasing their productivity (Hopp et al. 2009). Operations management scholars advocate that studies of system performance account for human behavior, which could alter the dynamics between operational variables and performance (Boudreau et al. 2003, Jouini et al. 2008). Empirical research that examines the interactions between human behavior and the design of operations systems can thus provide new insights for operations management theory.

Analytical models suggest that it is possible for dedicated queues to yield faster service times than pooled queues when servers have an incentive to invest in capacity so that they can increase their processing rates (Cachon and Zhang 2007, Gilbert and Weng 1998). Yet, to our knowledge, there has been limited field-based empirical study of the impact of differences in queue management systems on the speed of service. This is an important area for further investigation because without understanding both the human and system factors that impact the effectiveness of different queuing systems, operations management theory will be underdeveloped, leading to limited impact on practice.

To address this gap, we empirically test the impact of the structure of the queuing system on the throughput time of patients in an emergency department (ED) that switched from having a pooled queuing system to a dedicated queuing system. Our measure of throughput time was the patient's length of stay in the ED, a measure that started with arrival to the ED and ended with a bed request for admission to the hospital or the discharge of a patient to home or to an outside facility. The motivation behind the

ED's switch to a dedicated queuing system was to increase physicians' accountability for managing patient flow, with the goal of enabling the ED to handle the larger volume of patients that was predicted to occur. We find that, on average, the use of a dedicated queuing system—after controlling for individual patient, physician, and ED characteristics—decreased patients' lengths of stay by 10 percent. This represents a 32-minute reduction in length of stay—a meaningful time savings for the ED.

Operations management theory suggests a possible reason why the pooled queuing system had longer throughput times than the dedicated queuing system. Similar to workers in other service settings (Debo et al. 2008, Hasija et al. 2010, Tan and Netessine 2013), physicians in the dedicated queuing system had both the incentive and ability to make sure that their patients' care progressed efficiently, so that their patients in the waiting room could receive care sooner than they otherwise would have (Cachon and Zhang 2007, Gilbert and Weng 1998, Hopp et al. 2007, 2009). Thus, we find that the advantages of having visibility into one's own workload, in combination with both the ability to manage patient flow and an incentive to efficiently manage one's full workload, outweighed the variability-buffering benefit of a pooled queuing system. This paper makes a contribution to the literature on queue pooling because prior research has emphasized customer behaviors (e.g., jockeying) that reduce the process losses of dedicated queues, but fewer papers have empirically tested the impact of server behaviors on the performance of different types of queuing systems (Boudreau et al. 2003, Hopp et al. 2007, Jouini et al. 2008).

2. Prior Research and Hypotheses

2.1. Prior Research on Queue Management and Throughput 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). In this literature, queuing theorists have shown that shorter waiting times occur from pooling identical streams of customers into a single queue served by a pool of identically skilled servers of the same number as were in the original, dedicated systems (Kleinrock 1976, 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).

However, some analytical models have shown that the structure of customer demand and 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). For example, combining streams of customers who have different processing requirements can introduce inefficiencies which erode the benefits of pooling (Green and Nguyen 2001, Mandelbaum and Reiman 1998, Rothkopf and Rech 1987). Another factor that reduces the effectiveness of queue pooling is having strategic servers (Cachon and Zhang 2007, Debo et al. 2008, Hopp et al. 2007, 2009, Jouini et al. 2008). These strategic servers, as defined by Cachon and Zhang (2007) are able to adjust how much time they spend on tasks, what set of tasks to perform for a particular customer, in what order they carry out these tasks for that customer, and in what order to serve their set of customers.

Strategic servers can manipulate customer throughput time with these task management decisions when it benefits them to do so (Hopp et al. 2007, Link and Naveh 2006, Tan and Netessine 2013). For example, in the restaurant industry, Tan and Netessine (2013) find that wait staff adjust the services offered to customers so that customers spend less time in the restaurant when the workload is high. Similarly, Oliva and Sterman (2001) find that bank employees reduce the steps they go through to approve loans when their workloads are high, even though this erodes bank profitability. On the other hand, workers can also slow down their work pace when it benefits them to do so. Using analytical models, Debo and colleagues (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 throughput time for their customers. Similarly, Hasija and colleagues (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 throughput time is impacted by strategic servers' incentives and ability to do so.

One factor that may incentivize strategic servers to manipulate throughput times is the visibility into one's own workload. Some prior work has investigated this possibility. For example, in a laboratory experiment, Schultz and colleagues (1998) find that slower workers operating alongside faster co-workers on a serial production line speed up when inventory is low, because the low inventory state allows for easily visual monitoring of worker speeds as opposed to the high inventory state where it is not as readily perceived. Though this experiment was conducted in the context of interdependent work where workers experience a social pressure to not appear slow to co-workers, it highlights the impact that visibility into the workload may have on work speed and throughput times. In an empirical study of hospitals, KC and Terwiesch (2009) also find that workers alter their service rates depending on perceived workload.

This visibility into one's own workload is a characteristic that is, by design, present in dedicated queuing systems but absent in pooled queuing systems. In dedicated queuing systems, each server's workload is specified and visible. In contrast, pooled queuing systems provide little insight into one's own workload. Thus, strategic servers in dedicated queuing systems may be more incentivized to better

manage their workload, while those in pooled queuing systems may have insufficient incentives to do so.

2.2. Queue Management 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 presenting 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 an aspirin. The physician's choice can impact each individual patient's length of stay because of variance in the time required for the different options. In addition, the physician can impact throughput time by ensuring that others involved in patient care delivery, such as the nurse administering the aspirin or the radiology technician performing the scan, complete the tasks quickly. The physician can also impact 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 and an available physician. 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 greater visibility into his or her workload even before the patient is placed in an ED bed.

This increased visibility into one's workload in the dedicated queuing system is typically accompanied with a greater ability to manage the flow of one's patients because each physician manages his or her own bank of ED beds under a dedicated system. Thus, the accountability lies with the physician to free up the next bed and ensure that the next patient waiting in his or her queue is quickly placed into the newly available bed. In contrast, in a pooled queuing system, physicians typically rely on the triage nurse to manage the flow of patients into available beds for the entire ED, although the triage nurse has little to no control over the throughput time of patients. Thus, in a dedicated queuing system, physicians may have more ownership over the resources needed to process patients.

Prior operations management research suggests that this ownership and accountability might lead to lower throughput 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 a dedicated queuing system can produce shorter delivery times than a pooled queuing system because the suppliers have a greater incentive to invest in capacity so that they can quickly satisfy the buyer's orders, and thus receive a higher percentage of the orders. These studies suggest that when strategic servers have (a) visibility into their workload, (b) the ability to alter the flow rate through their system, and (c) an

incentive to more efficiently manage their workload, they are likely to decrease throughput times.

Through the context of an ED, we investigate how a switch from a pooled to a dedicated queuing system affected the behavior of strategic servers when the dedicated queuing system afforded a greater incentive and ability to efficiently manage customer waiting times. Specifically, we hypothesize that ED physicians may attain shorter throughput times when they work in an ED with a dedicated queuing system, due to the increased level of visibility into one's workload and the ability to better manage patient flow that is provided by the dedicated queuing system.

Hypothesis 1*: Patient length of stay will be shorter in the ED when physicians are working in a dedicated queuing system as opposed to a pooled queuing system.*

2.3. Social Loafing in the Emergency Department

When physicians work in an ED with a dedicated queuing system, they are more likely to attain shorter throughput times because the increased visibility into one's workload and the ability to better manage it may reduce social loafing. Social loafing theory, discussed by Latané and colleagues (1979), states that individuals may exert less effort in carrying out work when it is shared among many workers as opposed to only a few (Chidambaram and Tung 2005, Karau and Williams 1993, Latané et al. 1979). This behavioral phenomenon is distinct from—though similar to—moral hazard, which refers to the tendency for individuals to purposefully take more risk when the cost is borne by another individual (Arrow 1963, 1965, Pauly 1968, 1974, Spence and Zeckhauser 1971, Zeckhauser 1970). In the context of an ED, this suggests that physicians in a pooled queuing system may decrease their work rate towards the end of their shifts (i.e., carry out the same type of work but at a slower rate), since the patients remaining in the waiting room are perceived to be shared among multiple physicians. This would lead to a situation where physicians work slightly slower, even if they could feasibly be working faster.

In EDs with pooled queuing systems, physicians would be most incentivized to engage in social loafing towards the end of a shift, just before the buffer period of time when they are not assigned to care for any additional patients so that they can complete service for their current panel of patients. In many EDs, this buffer period constitutes the last one or two hours of a pre-determined shift. As the cutoff time marking the beginning of the buffer period nears, physicians may begin to work at a slower rate so that they are less likely to be assigned to care for the patients who are in the waiting room. In contrast, when working in an ED with a dedicated queuing system, no such incentive for social loafing would exist because each patient in the waiting room has already been assigned to a particular physician's queue.

To better understand the behavioral mechanism through which different queuing systems may impact a physician's throughput times, we explore the rate at which physicians provide service towards the end of a shift in a pooled versus a dedicated queuing system. Specifically, we hypothesize the following:

Hypothesis 2*: A physician's service rate towards the end of a shift will be will be greater 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 came from the emergency department (ED) of a 162-bed hospital in northern California. We selected this ED for study because in August 2008 it experienced an intervention—described in more detail below—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 used data from a time span before and after the intervention (March 2007 to July 2010) to test our hypotheses about the impact of queuing systems on throughput time in the ED. Depending on the time of day, the ED had an average of two to five physicians staffing 41 ED beds and up to nine hallway gurneys. This ED experienced an average five percent increase in patient census each year, from approximately 65,000 patients in 2007 to 76,000 patients in 2010. The average daily ED census was 178 patients in 2007 and 212 patients in 2010. This was a relatively large patient census in comparison to other EDs in the surrounding areas.

This ED, like many others, had a standardized patient flow process (Figure 1). Upon a patient's arrival, a registration clerk conducted a brief registration process. A triage nurse then 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 acuity) through 5 (lowest acuity). Higher acuity patients (ESI levels 1, 2, or 3) were typically treated in the main area (main ED). Lower acuity patients (ESI levels 4 or 5) were typically treated in the Rapid Care Area (RCA), unless main ED beds were available and the waiting room census was low or the patients arrived between 11pm and 7am when the RCA was closed.

---------- Insert Figure 1 About Here ----------

In this ED, the triage nurse assigned each patient to a specific attending physician, either upon 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 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. Physician shift times were determined in advance by the ED chief, and the ED scheduler assigned individual physicians to each of the pre-determined shift times. Physicians could change shifts on the hour between 5am and 11am, between 2pm and 5pm, and at 11pm or midnight. Between 7am and 11pm, there was usually one physician working in the RCA and four physicians

working in the main ED. During the overnight shift from 11pm to 7am, 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 their shifts by the ED scheduler. Physicians 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). 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. In order to allow physicians enough time to either complete a care plan or discharge the patients who had been assigned to them, physicians were assigned new patients only up until two hours before the end of their shifts. Patients arriving during the last two hours of a physician's shift were assigned to the oncoming physician, whose shift was scheduled such that it would begin two hours before the end of the preceding physician's shift. Because physician shifts were sufficiently staggered, this did not induce greater variation in system productivity.

3.2. Intervention: Change in the Physician Assignment System

In the main ED, an intervention—called the Physician Assignment System (PAS)—was implemented in August 2008. PAS effectively restructured the main ED from having a pooled queuing system to a dedicated queuing system.

Prior to the PAS intervention, higher acuity patients 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 triage nurse placed the next patient of highest acuity in this bed. Once a patient was placed in a bed, the triage nurse assigned a physician to the patient in a round robin fashion, which meant that physicians were assigned to patients in a set order regardless of their current workload. Once this assignment occurred, the assigned physician could see the 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 assigned by the triage nurse to an available ED bed.

The only exception to the round robin physician assignment 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. The round robin assignment policy was instituted to prevent physicians from unfairly selecting "easier" patients, and simultaneously made physician assignment to patients nearly random rather than due to a physician's 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 acuity patients. Thus, there was a limited amount of variation in patient intensity among the patients being assigned to the physicians staffing the main ED.

After the implementation of PAS, the triage nurse assigned each patient to a specific physician at the point of triage. In addition, each physician working in the main ED was designated an area of the main ED for which he or she was responsible, each of which comprised six main ED beds and two hallway gurneys. The one ED bed located in the resuscitation room remained unassigned to any particular physician and was reserved as buffer space for use by an ESI level 1 patient (i.e., a patient requiring immediate life-saving interventions). In addition, each main ED physician was designated two nurses with whom to care for patients, although each physician typically worked with more than 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 their shifts during which a relief nurse substituted in for the duration of the break. The round robin policy of assigning physicians to patients 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, physician assignment remained independent of a physician's speed of discharging patients.

After PAS, 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 and ready to be seen, but also those who were still in the waiting room. Thus, after PAS implementation, each physician worked under a dedicated queuing system with visibility into his or her dedicated queue of patients in the waiting room and a dedicated number of beds and nurses in the main ED. Compared to the pooled queuing system before the intervention, each physician now had an incentive and ability to manage patient flow in and out of their own beds. In addition, physicians were now unable to off-load patients still in the waiting room onto the oncoming physicians by working slower at the end of their shift.

In the RCA, the process used to assign patients to the RCA physician on duty did not change over the course of our study. A physician was assigned to a patient when he or she was called to be seen in an RCA examination room, not when they were in the waiting room. In other words, 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 used approximately 3.5 years of de-identified electronic medical record (EMR) data of all 243,248 patients treated in the ED from March 1, 2007 to July 31, 2010. The data extracted from the

EMR contained patient-level information including, but not limited to, the following: the patient's time of arrival and departure, length of stay, ESI level, attending physician, and disposition. We excluded patients with no attending physician or ESI level listed on their record, patients who left without being seen by a physician, and patients who had a length of stay of zero minutes or less. In addition, we excluded patients whose length of stay 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 excluded these observations from our dataset because their extended lengths of stay were typically driven by placement logistics rather than by physicians' levels of productivity. In addition, we excluded patients of ESI level 1 (i.e., patients needing resuscitation) and patients who died in the ED because their lengths of stay were likely to be driven by factors other than physician productivity. Lastly, we excluded trauma patients because these patients were primarily cared for by the hospital's trauma team, not a particular ED physician. Altogether, this resulted in an exclusion of 7,638 patients, which was 3.14 percent of the overall sample.

Using the final sample of 235,610 patients, we created a patient-level panel dataset that treats the physician as the panel variable. For the regression analysis, we limited our sample to the 231,081 patients who were 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).

To better understand workflow in the ED, we also conducted observations and interviews with ED staff. In the summer of 2012, the first author spent 86 hours shadowing ED physicians, consulting physicians, nurses, lab technicians, radiology technicians, pharmacists, transport service providers, and environmental service providers. This enabled her to observe the overall patient flow process, the triage process, the round robin assignment of physicians to patients, and how physicians and nurses provided care for multiple patients simultaneously. In addition, we conducted interviews with ED physicians, nursing staff, and the ED unit leadership about workflow in the ED before and after the implementation of PAS. We draw on transcripts from these interviews to shed deeper insight into our statistical findings.

3.4. Dependent Variables

Our primary dependent variable of interest was throughput time. To measure this, we used the patient's length of stay in the ED. For patients who were admitted to the hospital, this was defined as the time from a patient's arrival to the ED to the time the physician placed a bed request for admission to the hospital, thus excluding the time spent boarding in the ED and excluding any time spent in an inpatient unit. For patients who were discharged to home or to an outside facility, ED length of stay was defined as the time from a patient's arrival to the ED to his or her discharge from the ED. Our measure of throughput time, therefore, followed operations management convention by being the sum of waiting time and processing

time (Loch 1998). We log-transformed length of stay because its distribution was otherwise right-skewed.

3.5. Independent and Control Variables

3.5.1. Physician assignment intervention in main ED. The implementation of PAS marked the time at which the main ED transitioned from having a pooled queuing system to a dedicated queuing system. We captured this transition with a binary interaction term, $PAS \times main$, which was 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). Because it was unknown on exactly what date of August 2008 the PAS system had been implemented, and in order to account for an acclimation period, we omitted data from August 2008 in constructing the variable for PAS. We designated 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 accounted for several factors that may have affected our dependent variable of interest and may be correlated with our independent variables of interest. These included factors related to the patient's condition, the state of the ED, the physician's practice experience, and general time trends.

To account for the variation in length of stay due to the severity of a patient's condition, we controlled for the patient's acuteness and age. We accounted 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 was the best approximation we had for patient condition and severity because our dataset did not include patients' specific diagnoses (e.g., diagnosis-related groups (DRGs)). It was particularly 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 acuity levels and more patients presenting to the RCA were of lower acuity levels after the implementation of PAS.

To capture ED busyness and congestion, we controlled for the total number of physicians working during a given AM, PM, 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' lengths of stay that would arise from differences in structural elements of the ED, we controlled for the general time frame of the physician's shift (AM, PM, or overnight) and the location of the shift (main ED or RCA). We also controlled for whether an above-average productivity physician was present during the focal physician's shift, because previous studies have found that the presence of a fast worker on a shift creates social pressure which increases the productivity of the other employees (Mas and Moretti 2009). To create this binary variable, we first determined each physician's permanent productivity level by calculating the average length of stay for patients treated by the physician (Mas and Moretti 2009), adjusting for the full set of control variables

described in this section. We then categorized physicians whose permanent productivity level was greater than the 50th percentile as fast physicians (Mas and Moretti 2009). Because we wanted to control for the effect on length of stay that the presence of any *other* fast worker might have, regardless of the focal worker's productivity level, we constructed the variable such that it equaled 1 if there was at least one above-average productivity physician on the shift and 0 if there was no above-average productivity physician on the shift, not accounting for the productivity level of the focal physician.

To account for systematic differences arising from differences in physicians' experience working in this particular ED, we controlled for the number of shifts the physician had worked in this ED since the beginning of the dataset up until the point of each patient encounter. As we explain in more detail below, we also included physician fixed effects to account for other unobserved differences by physician.

Lastly, we accounted for time trends and related influences by adding dummy variables for each day of week and each month. For each patient encounter, we also accounted for the linear time trend by controlling for the number of months since the beginning of the dataset (March 2007).

3.6. Empirical Models

To test our hypotheses, we used linear regression models with physician fixed effects and clustered standard errors. Standard errors were clustered at the physician level 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 allowed us to control for unobservable individual physician effects that do not vary over time, such as level of motivation, innate ability, and practice routines. This is important to account for because they may significantly influence a physician's productivity level in ways that cannot be measured otherwise.

Specifically, we estimated the following two models:

$$\ln LOS_{ij} = \beta_o + \beta_1 PAS_{ij} + \beta_2 PAS_{ij} \times main_{ij} + \delta' \mathbf{X}_{ij} + \gamma' \mathbf{MD}_i + \varepsilon_{ij} \quad \text{(i)}$$

$$dischrate_{kj} = \alpha_0 + \alpha_1 PAS_{kj} + \alpha_2 PAS_{kj} \times penultimate_{kj} + \theta' \mathbf{X}_{kj} + \lambda' \mathbf{MD}_k + \omega_{kj} \quad \text{(ii)}$$

Model (i) is estimated at the patient level. Here, $\ln LOS_{ij}$ represents the logged number of minutes that patient *i* of physician *j* stayed in the ED; *PAS* indicates whether PAS had been implemented; *PAS* × *main* is an interaction term of whether PAS had been implemented and whether the shift was located in the main ED; **X** is a column vector of covariates; **MD** is a column vector of dummy variables that each represent a physician; prime (′) denotes transpose; β's and δ's represent vectors of coefficients; γ represents physician fixed effects; and ε is the time-varying error term not already captured by γ. Model (ii) is estimated at the physician-shift two-hour period level, where each observation represents the first, second, penultimate, or final two-hour period of a physician's shift. 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; *PAS* × *penultimate* is an interaction term of whether PAS had been implemented and whether this observation was of the penultimate two-hour period of a physician's shift; **X** is a column vector of covariates; **MD** is a column vector of dummy variables that each represent a physician; prime (′) denotes transpose; α's and θ's represent vectors of coefficients; λ represents physician fixed effects; and ω is the time-varying error term not already captured by λ. The column vector of covariates, **X**, includes all control variables described in the previous subsection, which includes month and day-of-week fixed effects, the linear time trend, the main effect for the shift location indicator variable (*main*) or shift period indicator variable (*penultimate*), respectively, among others. Table 1 provides summary definitions for all variables included in the models.

---------- Insert Table 1 About Here ----------

In addition to the standard assumptions of linear regression models, fixed effects models make two key assumptions, both of which were satisfied in our study. First was the assumption of strict exogeneity, which meant that the observation-specific error term of the model was uncorrelated with the covariates of both the observation and all other observations belonging to the same cluster (Wooldridge 2010). This was a plausible assumption to make in our context because (a) there was a low likelihood that patients with multiple visits would be treated by the same physician and (b) the patient error term was unlikely to be correlated with the covariates for other patients of the same physician. In addition, the unobservable random traits of patients that affected their patients' lengths of stay were not likely to be associated with the key independent variable of interest. Specifically, the round robin assignment of physicians to patients made it unlikely that the fastest physicians received the most complicated cases. In other words, physician assignment to patients was random and was not driven by physician speed or physician preference.

We chose to use fixed effects models rather than random effects models because we did not believe that the random effects assumption of zero correlation between the physician effect and the covariates (such as the number of shifts worked by the physician) would necessarily hold. By using fixed effects models, we were able to account for the unobserved traits of each physician that were associated with a patient's length of stay that were also correlated with the independent variables of interest. Accordingly, we conducted the Durbin-Wu-Hausman test, which rejected the random effects model in favor of the fixed effects model ($\chi^2 > 79.53$, $p < 0.001$).

To test Hypothesis 1, we estimated model (i). We used a difference-in-differences estimator to compare the difference in average throughput times in the main ED and the RCA before PAS implementation to the difference after PAS implementation. Because the physician assignment process did not change in the RCA, whereas the main ED moved from having a pooled to a dedicated queuing system, we considered 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 were 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 of stay, the RCA served as a reasonable control because our interviews with ED leadership and staff indicated that there were no other changes during the study period that affected only one part of the ED and not the other. Moreover, even though the main ED and the RCA treated patients of different acuteness, we expect that an operational change would bring about similar percentage change improvements, which is what we are examining by employing logged patient length of stay as the dependent variable.

To apply the difference-in-differences method, we first established the parallel trend assumption and calculated autocorrelation-consistent standard errors (Abadie 2005, Duflo 2001). We then estimated the effect of transitioning from a pooled to a dedicated queuing system on throughput times by examining the coefficient on the interaction term, $PAS \times main$. We predicted that this coefficient, β_2, would be negative and significant, suggesting that the dedicated queuing system was associated with shorter throughput times than the pooled queuing system. We also predicted that the coefficient on *PAS* would be non-significant, indicative of no effect of PAS in the RCA, which would mean that the only significant change affecting throughput times was the shift from a pooled to a dedicated queuing system in the main ED.

To test Hypothesis 2, we used the rate at which a physician discharged patients as our measure of service rate. We estimated model (ii) to examine whether the implementation of PAS in the main ED affected the discharge rate of patients towards the end of a physician's shift. Here, we limited the sample to patients seen in the main ED during the last four hours of a physician's shift and compared differences in the discharge rate during the penultimate two hours and final two hours of a physician's main ED shift before and after PAS implementation. This discharge rate was calculated at the physician-shift two-hour period level (i.e., for each of the two-hour periods of a physician's shift). Because the cutoff for being assigned new patients remained constant at two hours before the end of the shift and physicians were expected to complete their care for patients assigned to them prior to leaving, we considered the discharge rate during the final two hours of a physician's main ED shift as comprising the comparison group and the discharge rate during the penultimate two hours of a shift as comprising the treatment group.

If the reduction in throughput time were due to a reduction in social loafing when operating within a dedicated queuing system, we would expect that the discharge rate in the penultimate two hours of a physician's main ED shift would increase after the implementation of PAS. Specifically, we would expect that the difference in the discharge rates during the penultimate and final two hours of a physician's shift would be greater after PAS compared to the difference before PAS. In other words, we predicted that the coefficient on the interaction term, $PAS \times penultimate$, would be positive and significant. Here, we predicted that the coefficient on *PAS* would also be positive and significant, since the earlier assignment

of physicians to patients is likely to increase the total number of patients seen during a shift, thus increasing the number of patients needing to be discharged by the physician by the end of the shift.

To better understand our main findings and consider possible alternate explanations, we conducted several additional analyses. First, we examined whether the hypothesized change in throughput time was driven by a change in ED wait time, a change in ED care delivery time, or both. In addition, we examined whether there was any differential change in ED boarding time after the PAS intervention. ED wait time was calculated as the time elapsed between when the patient arrived to the ED and when the physician began caring for the patient. ED care delivery time was calculated as the time elapsed between when the physician began caring for the patient and when the patient was discharged. ED boarding time was calculated only for patients who were admitted to an inpatient unit, and was equal to the time elapsed from when the physician placed a bed request for admission to the hospital and when the patient left the ED. ED wait time, care delivery time, and boarding time were all measured at the patient level. They were captured in minutes and subsequently log-transformed due to their right-skewed distributions. For this set of analyses, we used the same specification as model (i) but, but substituted logged length of stay with logged ED wait time, care delivery time, and boarding time, respectively, as the dependent variable. If the reduction in overall length of stay was due to a reduction in social loafing under the dedicated system, we would expect the reduction in overall length of stay to be driven by a reduction in ED care delivery time rather than a reduction in ED wait time, though both may be statistically significant because the care delivery times affect the wait times of incoming patients. Thus, with logged ED wait time and logged ED care delivery time as the dependent variable, respectively, we predicted that the coefficient on $PAS \times main$ would be negative and significant. However, because ED boarding time is determined by the inpatient unit's capacity to admit a new patient, we expect that there is no significant change in this measure. Thus, we predicted that the coefficient on $PAS \times main$ would not be statistically significant.

We also considered two competing explanations that could account for the decrease in throughput time post-PAS besides a reduction in social loafing. First, patients might have experienced shorter lengths of stay in the ED because physicians "cut corners" by stinting on care (Oliva and Sterman 2001). We assessed this possibility by estimating model (i) with two different dependent variables, both measured at the patient level: whether labs were ordered for a patient and whether x-rays were ordered for a patient. Data on whether labs or x-rays were ordered for a patient were obtained directly from the hospital's EMR system. For each of these variables, we estimated model (i) as a logistic regression because each of the dependent variables was a binary indicator variable. Second, we considered whether the decrease in length of stay stemmed from physicians shifting their work onto other clinicians. In the context of the ED, the most plausible scenario would be that of ED physicians having more patients admitted to the hospital, so that the patients appear to stay in the ED for a shorter period of time. We examined this possibility by

estimating model (i) with admission to the hospital as the dependent variable. Data for whether the patient was admitted to the hospital came from the EMR system and was measured at the patient level. Again, we estimated a logistic regression because admission to the hospital was a binary dependent variable.

Lastly, we examined a potential unintended consequence of the quality of care in the ED becoming worse after the implementation of PAS in the main ED. As a proxy for quality, we used the patient's mortality in the ED, and estimated a logistic regression with this binary indicator as the dependent variable. For this analysis, we included a subset of the patient-level observations that were previously excluded—specifically patients of ESI level 1, patients who died in the ED, and trauma patients.

4. Results

4.1. Descriptive Statistics

Table 2 presents means, standard deviations, and correlations for all continuous variables and percentages for all categorical or binary variables included in the empirical model. Descriptive statistics are presented for the overall sample, and then separately for the pre-PAS period and the post-PAS period. None of the correlations between variables that were used in the same regression model had levels close to or higher than 0.80, which would have triggered concerns about multicollinearity. In addition, the largest variance inflation factor (VIF) is 6.89 and the mean VIF is 2.35, both of which fall below the conventional threshold of 10 (Wooldridge 2012).

We find that the average length of stay for a patient seen in this ED was 199 minutes (*s.d.* = 199). Specifically, the average length of stay was 248 minutes for a patient seen in the main ED (*s.d.* = 220) and 96 minutes for a patient seen in the RCA (*s.d.* = 73). The average patient was 37.8 years old (*s.d.* = 24.4). An average ED physician had almost 13 years of experience working as a physician (*s.d.* = 7.9), and worked on average 255 shifts (*s.d.* = 163) in this ED between March 1, 2007 and July 31, 2010. On average, physicians were on shift for almost 10 hours (*s.d.* = 1.4). There were approximately five physicians (*s.d.* = 1.2) staffing the entire ED during a given eight-hour period (i.e., AM shift, PM shift, overnight shift), with up to nine physicians when counting the total number of physicians that were present for any portion of the eight-hour period (due to staggered start times). On average, 35 patients (*s.d.* = 12) were in this ED at any given time.

---------- Insert Table 2 About Here ----------

Approximately eight percent of patients were of ESI level 2, 51 percent were of ESI level 3, 39 percent were of ESI level 4, and one percent were of ESI level 5. Of patients seen in the main ED, 74 percent were of ESI level 3. Patient arrivals were approximately uniformly distributed across the months of the year. Of the days of the week, arrivals peaked on Mondays and were also relatively high on Saturdays and Sundays. Approximately half of all patients arrived during the PM shift and another 16

percent arrived during the overnight shift. Sixty-eight percent of patients were seen in the main ED.

As expected, patients' average length of stay differed significantly by their acuteness. Table 3 presents the means and standard deviations of patients' average length of stay by ESI level, as well as the frequencies of each ESI level. We find that, for patients of ESI levels 2 to 5, the relationship between length of stay and ESI level is a generally monotonically increasing one, where patients of a higher acuteness (e.g., ESI level 2) had a longer length of stay and those of a lower acuteness (e.g., ESI level 5) had a shorter length of stay. We account for the non-linearity of this relationship by adjusting for patient acuteness using a dummy variable for each ESI level.

---------- Insert Table 3 About Here ----------

4.2. Physician Assignment System Implementation in Main ED

Both the quantitative and qualitative data suggested that PAS was implemented as described, though not without challenges. One of the ED physicians 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, suggested that physicians largely abided by the physician assignment processes outlined by PAS.

However, there were rare situations during which PAS and the associated round robin assignment system were violated. Physicians working in the main ED could use their discretion to bypass the round robin assignment determined by PAS when another physician had an exceptionally time-consuming workload of ESI level 1 patients. One physician described: "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 corroborated our understanding of the round robin assignment system, in which other physicians could voluntarily take on the next patient assigned to a physician caring for an ESI level 1 patient.

In our EMR data, we found further support for the general adherence to the round robin assignment system. In particular, patient demographics across physicians were well balanced, suggesting that it was not the case that certain physicians were likely to be assigned particular types of patients. In particular, there was little variation in the acuteness of patients who were assigned to the different physicians working in the main ED. Furthermore, on average there were 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 that the workload across physicians remained relatively balanced, thus allowing the physicians to feasibly adhere to the round robin system of assigning physicians to patients.

4.3. Base Results

We estimated model (i) to assess the impact of having a pooled queuing system (versus a dedicated queuing system) on throughput time. The results of our analysis are summarized in model (1) of Table 4.

---------- Insert Table 4 About Here ----------

Table 4 presents a fixed effects model that captures the effect of moving from a pooled queuing system to a dedicated queuing system. We find that the difference in throughput times between the main ED and the RCA is greater prior to PAS implementation. Once the main ED adopted a dedicated queuing system, this difference in throughput times reduced. This difference-in-differences is captured by the coefficient on the interaction term, $PAS \times main$ ($\beta_2 = -0.10$, $p < 0.001$), and indicates that the transition from a pooled queuing system to a dedicated queuing system is associated with a highly significant reduction in patients' lengths of stay. This 10.05 percent decrease in a patient's length of stay in the main ED after the implementation of PAS corresponds to a decrease of 32 minutes for an average patient of ESI level 3 seen by an average physician in the main ED. In other words, the average patient's length of stay was significantly longer in the pooled queuing system than in the dedicated queuing system. The coefficient on *PAS* is not statistically significant at conventional levels ($\beta_1 = -0.03$, $p \approx 0.10$), suggesting that there was no significant change in patients' lengths of stay in the RCA. These results offer strong support for our main hypothesis, which predicted that, in our setting, pooled queuing systems are associated with longer throughput times compared to dedicated queuing systems.

Next, to better understand the mechanism through which dedicated queuing systems may impact throughput times, we estimated model (ii). We examined whether the implementation of PAS in the main ED affected the discharge rate immediately before and after the physician assignment cutoff time at two hours before the end of a physician's shift. These results are summarized in model (1) of Table 5.

---------- Insert Table 5 About Here ----------

Model (1) of Table 5 presents a fixed effects model estimated at the physician-shift two-hour period level. As predicted, we find that the discharge rate in the main ED exhibits a slight increase overall after PAS implementation, which is consistent with the finding that physicians are assigned slightly more patients during a shift after PAS implementation. Specifically, the discharge rate in the final two hours of a physician's main ED shift increases by 0.1 patients per hour ($\alpha_1 = 0.12$, $p < 0.001$), and it increases by an additional 0.05 patients per hour ($\alpha_2 = 0.05$, $p < 0.03$) in the penultimate two hours of a physician's shift after PAS implementation. This is consistent with Hypothesis 2 that a dedicated queuing system would discourage social loafing due to the greater visibility into individually assigned responsibilities that it affords. This incentive to better manage one's full workload, in combination with the increased ability to manage patient flow directly, results in an increase in discharge rates after PAS implementation in the penultimate two hours of a physician's shift relative to the final two hours of the shift. The 0.05 increase

in the difference of discharge rates in the main ED in the penultimate versus final two hours of a physician's shift suggests that the reduction in throughput times after PAS implementation may be attributable to the reduction in social loafing when operating within a dedicated queuing system.

To examine whether the reduction in throughput time was driven by a change in ED wait time, a change in ED care delivery time, or both, we repeated the estimation of model (i) at the patient level with logged ED wait time and logged ED care delivery time as the dependent variable, respectively. These results are presented in models (2) and (3) of Table 4. In model (2), which takes logged ED wait time as the dependent variable, we find an 8.53 percent reduction in the difference in ED wait times between the main ED and the RCA after PAS implementation ($\beta_2 = -0.09$, $p < 0.01$). In model (3), with logged ED care delivery time as the dependent variable, we find an even greater reduction in the difference in ED care delivery times between the main ED and the RCA after PAS implementation ($\beta_2 = -0.17$, $p < 0.001$). This indicates that the decrease in ED wait time and ED care delivery time are both contributory factors to the decrease in patient length of stay, but that the decrease in ED care delivery time after the implementation of PAS is a larger driver of this overall reduction in throughput time.

In addition, we examined whether the implementation of PAS affected patients' boarding times, during which patients who are to be admitted to the hospital wait to be transferred out of the ED to an inpatient unit. These results are presented in model (4) of Table 4. Consistent with our prediction, we find that there is no significant change in ED boarding times after PAS implementation ($\beta_2 = -0.24$, $p \approx 0.13$). This suggests that boarding times are indeed minimally affected by ED physicians' productivity levels, and primarily determined by the inpatient unit's capacity to admit a new patient.

4.4. Consideration of Alternate Explanations and Unintended Consequences

Though our finding of reduced throughput times in a dedicated queuing system is consistent with a reduction in social loafing during the penultimate two hours of a physician's shift, we consider alternate explanations that could also be consistent with our main finding. We also explore the possibility of unintended consequences arising as a consequence of implementing a dedicated queuing system.

4.4.1. Testing for changes in the provision of care. First, one possibility is that physicians change their practice behaviors after the implementation of PAS in a way that they stint on care because of the increased pressure to care for all of the patients in their dedicated queue. If fewer services are being provided for patients, they may be staying in the ED for a shorter amount of time. For example, if a patient who would have otherwise gotten an x-ray does not receive an x-ray, she would likely stay in the ED for a shorter duration because (a) she would not need to wait for the x-ray machine to become available, (b) she would not need to have the x-ray taken, and (c) she would not need to wait for the radiologist to read the films. If physicians were stinting on care, we would be mistaken to assume that the reduced length of stay stems from a reduction in social loafing.

We do not find strong evidence of stinting on care after the transition to a dedicated queuing system in the main ED. In model (1) of Table 6, we examine the change in a patient's likelihood of having a lab test ordered. We find that the coefficient for *PAS* × *main* is not significant, suggesting that the difference in the likelihood of having a lab test ordered for a patient in the main ED and the RCA did not change significantly after PAS implementation. However, we find that the coefficient for the main effect, *PAS*, is positive and significant, which suggests that the likelihood of having a lab test ordered for a patient in the RCA increased after PAS was implemented. This finding may stem from the fact that the network of hospitals to which this emergency department belonged instituted a network-wide initiative to improve sepsis identification in late November of 2008. This initiative—which was implemented in both the main ED and the RCA—defined as the new standard of care the administration of a certain kind of lab test when patients present with a predetermined set of symptoms. This is likely to have affected care in the main ED and the RCA in a similar fashion. Yet, because our data does not capture the count of lab tests ordered but only an indicator for whether at least one test was ordered or not, and because the pre-PAS likelihood of receiving a lab test was eight percent in the RCA and 64 percent in the main ED, it is likely to have contributed to a greater binary increase in the RCA than in the main ED. Indeed, when we limit the time frame of our analyses to the three months before and after PAS implementation—before the sepsis initiative—we find that there are no significant changes in the likelihood of receiving a lab test after PAS in both the RCA ($\beta_1 = -0.05$, $p \approx 0.69$) and the main ED ($\beta_2 = 0.01$, $p \approx 0.96$).

Similarly, in model (2) of Table 6, we do not find a significant change in a patient's likelihood of having an x-ray ordered. Here, neither the coefficient on the interaction term ($\beta_2 = -0.03$, $p \approx 0.48$) nor the coefficient on *PAS* ($\beta_1 = 0.03$, $p \approx 0.40$) is statistically significant, suggesting that there was no meaningful change in a patient's likelihood of receiving an x-ray before and after the implementation of PAS. In combination, these results suggest that physicians are not systematically stinting on care as a result of PAS implementation as measured by the likelihood of lab tests and x-rays being ordered.

---------- Insert Table 6 About Here ----------

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' length of stay 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 length of stay in the ED may appear to be shorter than it would be otherwise.

We examine this possibility by estimating a similar logistic regression with whether a patient is admitted to the hospital as the dependent variable. If shorter lengths of stay are attained by having more patients from the main ED admitted to the hospital after PAS, we would expect the coefficient on *PAS* × *main* to be positive and significant. In contrast, as demonstrated in model (3) of Table 6, 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 (β_2 = -0.17, $p \approx 0.25$). Thus, we do not find strong evidence of an increased likelihood of admitting patients to the hospital after the transition to a dedicated queuing system in the main ED.

4.4.3. Testing for changes in the quality of care. Next, we consider a potential unintended consequence of this transition from a pooled queuing system to a dedicated queuing system in the main ED. We assess whether physicians change their practice behaviors after the implementation of PAS in a way that they provide a lower quality of care. If physicians are providing lower quality care such that more patients are dying in the ED, this truncating effect on length of stay may result in a decrease in the average patient's throughput time in the ED. This is both a possible alternate explanation for the decrease in length of stay and a potential unintended consequence of the intervention.

In our analysis, we do not find evidence of lower quality of care as measure by mortality in the ED. These results are presented in model (4) of Table 6. Due to 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 (β_2 = -0.79, $p \approx 0.03$). This suggests that the quality of care, as measured by patient mortality in the ED, may have 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.

Another proxy for quality that is often measured in health care settings is patient readmission. To comply with the Institutional Review Board requirements, our data lacks the patient identifiers that would be required to assess readmission at the patient level. Therefore, our data does not allow for an examination of changes in readmission rates at the patient level. However, we were able to obtain this data at the ED level for every month from March 2007 to July 2010. Although this data limits the rigor with which we are able to conduct the analysis, we find that the 48-hour and 7-day readmission rates do not change significantly before and after PAS implementation. At the ED level, the 48-hour readmission rate decreases slightly in magnitude from 2.92 percent to 2.85 percent, although this difference is not statistically significant. The 7-day readmission rate also exhibits a non-statistically significant decrease from 6.56 percent before PAS implementation to 6.53 percent after PAS implementation.

4.4.4. Testing for potential impact on the duration of a physician's shift. Lastly, we consider the potential impact of PAS on the duration of a physician's shift. Though this does not directly address why having a dedicated queuing system may decrease patients' lengths of stay, it is important to consider in order to assess the feasibility of implementing such a system at other EDs. If having a dedicated queuing system results in physicians staying longer on their shift 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 model (ii) but with the duration of a physician's shift as the dependent variable and at the physician-shift level (as opposed to the physician-shift two-hour period level). We estimate this regression at the physician-shift level because the dependent variable (i.e., shift duration) is calculated at this level. If physicians were working longer hours as a result of PAS implementation, we would expect to see a positive and significant coefficient on the interaction term, $PAS \times main$. However, we do not find evidence in support of this. As is summarized in model (2) of Table 5, we find that there is no significant difference between the duration of a shift in the main ED and the RCA before and after PAS implementation ($\alpha_2 = -0.10$, $p \approx 0.26$).

4.5. Specification Tests

To examine the robustness of our findings, we tested a variety of other specifications in addition to the reported models. First, we used a limited model specification that included only patient ESI levels as control variables. We retained patient ESI levels because the average acuteness of patients arriving in the main ED became higher over time while that of patients arriving in the RCA became lower over time. We find that the base result remains robust to this limited model specification ($\beta_2 = -0.07$, $p < 0.001$), though the magnitude of the effect decreases slightly from 10.05 percent to 7.27 percent.

Next, we limit our sample to those patients seen in the main ED and conduct a pre-post analysis, comparing the average length of stay of patients before and after the implementation of PAS. We find that our main findings are very robust to this alternate specification, where the implementation of PAS is associated with a 12.05 percent decrease in length of stay in the main ED ($\beta_2 = -0.12$, $p < 0.001$).

Although our interviews with ED staff suggested that there were no other interventions besides PAS that were applied to only the main ED or the RCA during the study period (March 1, 2007 to July 31, 2010), we applied our analyses to shorter time frames around PAS implementation to nullify the possibility of other effects. When we limited the time frame to three months, seven months, 12 months, 15 months, and 18 months before and after the intervention, we found that our base results remained robust to these shorter time frames ($\beta_2 < -0.07$, $p < 0.002$).

We also examined two alternate specifications of the variable that was used in our models to indicate the presence of a fast physician on a focal physician's shift. In one, we categorized physicians with permanent productivity levels above the 75th percentile as fast physicians. In the other, we used a continuous variable that was a count of the number of fast physicians on a shift. Our base results remained robust to these alternate specifications ($\beta_2 < -0.10$, $p < 0.001$).

In addition, our results do not appear to be due to differences in patient care delivered in the two areas of the ED. To examine this, we assessed whether the transition from a pooled system to a dedicated system differentially affected length of stay depending on the location of a patient's ED care. To conduct

this analysis, we used the same empirical model as specified above, but limited the sample to patients of ESI levels 4 and 5, and with each independent variable of interest interacted with ESI level 5. We limited the sample to these patients because they constituted the group of patients who were potentially seen in both areas of the ED (because all ESI level 4 and 5 patients were seen in the main ED after 11pm). This analysis suggests that there are no differential effects by the location of a patient's ED care ($p > 0.12$).

Furthermore, we examined whether the base results are sensitive to heterogeneity in patient acuteness. In other words, we examined whether the transition from having a pooled queuing system to a dedicated queuing system had 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 explored this possibility by limiting the sample to patients of ESI levels 2 and 3, and interacted each independent variable of interest with ESI level 3. For this analysis, we limited the sample to patients of these two ESI levels because they exhibited two different groups with relatively longer (ESI level 2) and shorter (ESI level 3) average lengths of stay. This analysis suggests that patients of higher acuteness (ESI level 2) were likely to experience a slightly greater decrease in length of stay after the implementation of PAS compared to patients of a relatively lower acuteness (ESI level 3). While 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 acuity patients (ESI level 2) within each physician's dedicated queue.

We also repeated our analyses using different exclusion criteria in constructing our sample. First, we included all observations that had previously been excluded as outliers (i.e., patients with a length of stay greater than 48 hours). Then, we amended our outlier cutoff to exclude observations with a length of stay greater than one day (24 hours) and the average duration of one shift (9.4 hours), respectively. In addition, we included ESI level 1 patients in our sample, excluded patients arriving by ambulance, and excluded patients presenting with a psychological condition, respectively. All coefficients of interest and their corresponding significance levels remained robust to these alternate specifications ($\beta_2 < -0.08$, $p < 0.001$).

In addition, we repeated our analyses using an alternate measure of patient length of stay. For all patients regardless of ED disposition, we defined length of stay as the time from a patient's arrival to the ED to his or her discharge from the ED. In other words, this included boarding time in a patient's length of stay, which had previously been excluded from the length of stay of patients who were subsequently admitted to the hospital. Using the log-transformed version of this alternate measure of length of stay, we found that that the transition from having pooled to dedicated queues was associated with a 13.80 percent decrease in patient length of stay ($p > 0.001$).

Lastly, we used hierarchical linear models, which specify random effects rather than fixed effects at the physician level. This specification test was conducted in order to test each of our hypotheses with greater efficiency gains. Here, we used three levels for our multilevel analyses: patient-level, physician

shift-level, and physician-level. The effect of transitioning from having a pooled queuing system to a dedicated queuing system remained robust to this model specification (β_2 = -0.10, $p < 0.001$).

5. Discussion and Conclusions

Using 3.5 years of data from a hospital's ED, we found that patients experienced shorter lengths of stay when physicians were working in a dedicated queuing system as opposed to a pooled queuing system. Our findings were consistent with the analytical models that suggest it is possible for dedicated queues to yield faster service times than pooled queues when strategic servers are able and incentivized to manage their capacity to achieve the goal of increasing processing rates (Cachon and Zhang 2007, Gilbert and Weng 1998). We suggest that the improved performance in a dedicated queuing system stems from a reduction in social loafing, which was a behavior that was triggered in the pooled queuing system by having patients who could be processed by any physician rather than by a specific physician. We considered, but did not find empirical support for, alternate explanations for this reduction in patient length of stay, such as changes in the provision of care or lower quality care in the ED.

We found evidence of a reduction in social loafing towards the end of a physician's shift, when the physician nears the cutoff period after which no additional patients are assigned. In our examination of physicians' discharge rates immediately before and after this cutoff period, we found that the discharge rate in the penultimate two-hour period of a physician's shift differentially increased after the transition to a dedicated queuing system (Table 5, model 1). This suggests that the benefits associated with the greater visibility into one's workload and an increased ability for physicians for manage patient flow under a dedicated queuing system may outweigh the variability-buffering benefits of a pooled queuing system. In describing how the implementation of PAS increased 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 [triage] nurse onto me to manage my patients."

In the context of our study setting, we find it particularly important to consider the significance of the effect sizes found in our analyses. For example, we find that moving from a pooled queuing system to a dedicated queuing system is associated with a 10.05 percent decrease in a patient's length of stay (Table 4, model 1). For an average patient of ESI level 3 seen in the main ED by an average physician, this corresponds to a decrease in length of stay of 32 minutes. 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 107 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. For example, in our study setting, we found that the main ED's transition from a pooled to a dedicated queuing system was associated with a significant *increase* in care delivery time in the RCA where the intervention did not take place (Table 4, model 3). This may have been an unintended negative consequence that stemmed from focusing managerial attention on process changes in the main ED at the expense of work in the RCA. This observation is consistent with similar findings in the process improvement literature, which suggests improvement efforts may have negative consequences in the short term because managerial attention is diverted from regular production (Repenning and Sterman 2002).

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 throughput time. We find that when servers are strategic such that they have the ability to manage their capacity, visibility into their workload, and an incentive to create fast throughput times, dedicated queuing systems are associated with *shorter* throughput times than pooled queuing systems. Our finding illustrates the importance of accounting for the interaction between human behavior and queue design when predicting system performance (Boudreau et al. 2003, Jouini et al. 2008). When queuing theory does not take strategic behavior into account, it instead suggests that pooling queues should result in shorter throughput times (Jouini et al. 2008). Thus, our paper provides empirical support for prior analytical models that predicted that human behaviors could reduce the positive benefit of using a common pool (Cachon and Zhang 2007, Gilbert and Weng 1998, Hopp et al. 2007, Jouini et al. 2008, Wang et al. 2010).

Our paper demonstrates how individual tendencies to engage in social loafing may undermine the potential benefits of pooling. As hypothesized, our analyses suggest that queue pooling results in longer throughput times because physicians engage in social loafing if it is possible for another physician to assume responsibility for the next patient. This result is similar to, but different from, Chan's finding (2013) that ED physicians worked slower when they were assigned patients by a triage nurse than when physicians—collectively as a group—assigned patients to physicians. Chan asserts that this "foot-dragging" behavior occurred in the nurse-managed system because physicians would delay discharges or

distort censuses to overstate their true workload to the nurse in hopes of avoiding being assigned another patient. The findings in Debo and colleagues' (2008) study are also driven by misleading behaviors. The customer's limited knowledge about the true level of service that should be provided enables workers to increase throughput time to gain higher payments per customer during periods of low demand.

In contrast, we highlight a different underlying mechanism for the improvement in throughput times. In our study, a triage nurse used a round robin assignment system to assign physicians both before and after the intervention. Thus, physicians were not working slower to overstate their workloads. Instead, we find that reducing social loafing by making a single physician accountable for efficiently managing the throughput time of a group of waiting patients leads to a reduction in throughput time. Thus, our results are similar to the predictions of Gilbert and Weng (1998) and Cachon and Zhang (2007), which find that dedicated queues can result in faster performance if they are combined with ability and incentives to increase capacity.

An additional contribution is that we quantify the positive effect of using a dedicated queuing system to increase servers' accountability for managing customers' throughput times in an actual working environment. In our context of a hospital ED, we find that having a dedicated queuing system, as opposed to a pooled one, is associated with a 10 percent increase in patients' length of stay. Thus, we are able to add evidence to the debate that lean manufacturing's practice of increasing workload visibility and accountability through the assignment of a specific person to a particular stream of work outperforms the pooling benefits of allowing any available worker from a pool of employees to respond to work requests (Spear and Bowen 1999).

5.2. Limitations and Future Research

This study has limitations, and its results should be interpreted accordingly. First, we note the threat of omitted variable bias that is common to many empirical models. While it would have been helpful to include more patient characteristics in our model, such as patient diagnosis or medical comorbidities, these data were protected information and not available for use. However, this is not an important threat because physicians were randomly assigned to patients rather than by preference. This is supported by the fact that the average ESI level of patients seen by each physician was 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 limited to one hospital's ED and its response to one intervention. This limits the generalizability of our findings, though we believe our findings have strong theoretical underpinnings. Nevertheless, we welcome future research on these effects and mechanisms in different empirical contexts for further substantiation.

Third, while it is beyond the scope of this paper to thoroughly examine the effects of increased productivity, decreased throughput times, or dedicated queuing systems on various quality measures, this

is an important element to consider in future research. In this paper, we were limited to examining quality via proxies such as patient mortality in the ED. We were not able to extend our analyses to include the effects on quality as measured by clinical outcomes due to the protected nature of these data. Future research should also consider how dedicated queuing systems may affect patient and physician satisfaction, since changes in throughput times may to be associated with perceptions of fairness and the general satisfaction of both parties. These data were not available at the time of this study.

Lastly, implementing a dedicated queuing system is merely one way to try to attain the goal of reducing patient length of stay in EDs. Future research should consider what other mechanisms may exist to try to reduce throughput times in similar settings, such as managerial incentives or interventions that leverage social pressure (Chan 2013). For example, do physicians increase their work rates when provided information about each other's average throughput times? It may be possible to use a combination of interventions so that EDs can capture the benefits of pooling while simultaneously avoiding social loafing.

5.3. Practical Implications and Conclusions

Our study has important implications for workplace managers and health care policy makers. Managers of work settings with strategic servers should design work systems to mitigate behaviors that benefit the employee to the detriment of the customers or the organization. We find that one mechanism is to give strategic servers greater ability and incentive to manage the workflow and to make the workload constant regardless of work pace, which removes the benefit of engaging in social loafing by slowing down. Our findings suggest that, in workplaces where servers are strategic, the potential negative effects of designing systems with pooled queues must be carefully considered. This has implications for designing and managing staffing structures and workflows, particularly in the context of service delivery organizations.

Specifically in the context of health care, our findings suggest that EDs may benefit from implementing dedicated queuing systems in which physicians are assigned to patients immediately upon patient arrival following triage. To our knowledge, this is not currently in place at most EDs; instead, most EDs employ a pooled queuing system that assigns physicians to patients once they are placed in an ED bed. Thus, the potential for improvement is significant.

While results may differ across different settings, the mechanisms through which changes in throughput time occurred may help shed light on cost savings predictions in other contexts. Our findings are especially timely and could have significant implications for health care 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).

References

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

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

Arrow, K. J. 1963. Uncertainty and the welfare economics of medical care. *Amer. Econom. Rev.* **53**(5) 941–973.

Arrow, K. J. 1965. *Aspects of the Theory of Risk-Bearing*. Yrjö Jahnssonin Säätiö, Helsinki, Finland.

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

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

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

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

Chan, D. C. 2013. Teamwork and moral hazard among emergency department physicians. Working paper, MIT, Cambridge, MA.

Chidambaram, L., L. L. Tung. 2005. Is out of sight, out of mind? An empirical study of social loafing in technology-supported groups. *Inform. Systems Res.* **16**(2) 149–168.

Debo, L. G., L. B. Toktay, L. N. Van Wassenhove. 2008. Queuing for expert services. *Management Sci.* **54**(8) 1497–1512.

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, G. D. 1979. Note - Effects of centralization on expected costs in a multi-location newsboy problem. *Management Sci.* **25**(5) 498–501.

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

Gilbert, S. M., Z. K. Weng. 1998. Incentive effects favor nonconsolidating queues in a service system: The principal-agent perspective. *Manage. Sci.* **44**(12) 1662–1669.

Green, L. V, V. Nguyen. 2001. Strategies for cutting hospital beds: The impact on patient service. *Health Services Res.* **36**(2) 421–42.

H.R. 3590 - 111th Congress. 2010. *Patient Protection and Affordable Care Act*. United States.

Hasija, S., E. Pinker, R. A. Shumsky. 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, W. J., S. M. R. Iravani, F. Liu. 2009. Managing white-collar work: An operations-oriented survey. *Production Oper. Management* **18**(1) 1–32.

Hopp, W. J., S. M. R. Iravani, G. Y. Yuen. 2007. Operations systems with discretionary task completion. *Management Sci.* **53**(1) 61–77.

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

Karau, S., K. Williams. 1993. Social loafing: A meta-analytic review and theoretical integration. *J. Personality Soc. Psych.* **65**(4) 681–706.

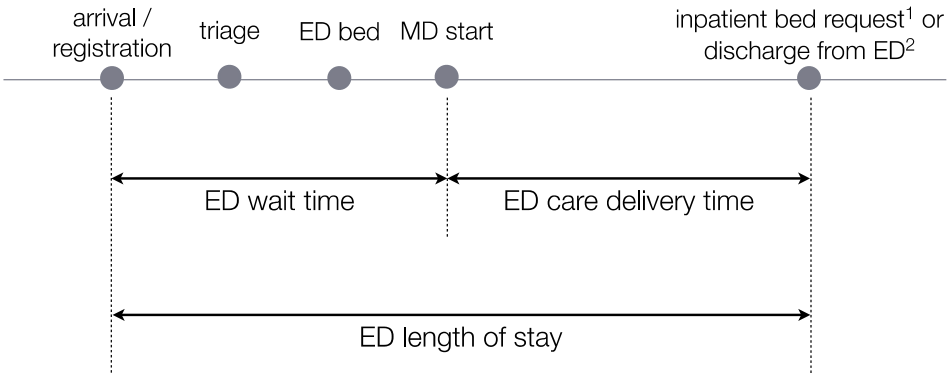
KC, D. S., C. Terwiesch. 2009. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Sci.* **55**(9) 1486–1498.

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

Latané, B., K. Williams, S. Harkins. 1979. Many hands make light the work: The causes and consequences of social loafing. *J. Personality Soc. Psych.* **37**(6) 822–832.

Link, S., E. Naveh. 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., M. I. Reiman. 1998. On pooling in queueing networks. *Management Sci.* **44**(7) 971–981.
Mas, A., E. Moretti. 2009. Peers at work. *Amer. Econom. Rev.* **99**(1) 112–145.
Oliva, R., J. D. Sterman. 2001. Cutting corners and working overtime: Quality erosion in the service industry. *Management Sci.* **47**(7) 894–914.
Pauly, M. V. 1968. The economics of moral hazard: Comment. *Amer. Econom. Rev.* **58** 531–536.
Pauly, M. V. 1974. Overinsurance and public provision of insurance: The roles of moral hazard and adverse selection. *Quart. J. Econom.* **88**(1) 44.
Repenning, N. P., J. D. Sterman. 2002. Capability traps and self-confirming attribution errors in the dynamics of process improvement. *Admin. Sci. Quart.* **47**(2) 265–295.
Rothkopf, M. H., P. Rech. 1987. Perspectives on queues: Combining queues is not always beneficial. *Oper. Res.* **35**(6) 906–909.
Schultz, K. L., D. C. Juran, J. W. Boudreau, J. O. McClain, L. J. Thomas. 1998. Modeling and worker motivation in JIT production systems. *Management Sci.* **44**(12) 1595–1607.
Spear, S., H. K. Bowen. 1999. Decoding the DNA of the Toyota production system. *Harvard Bus. Rev.* **77**(5) 96–106.
Spence, M., R. J. Zeckhauser. 1971. Insurance, information, and individual action. *Amer. Econom. Rev.* **61**(2) 380–387.
Tan, T., S. Netessine. 2013. When does the devil make work? An empirical study of the impact of workload on worker productivity. Working paper, INSEAD, Fontainebleu, France.
van Dijk, N. M., E. van der Sluis. 2008. To pool or not to pool in call centers. *Production Oper. Management* **17**(3) 296–305.
van Dijk, N. M., E. van der Sluis. 2009. Pooling is not the answer. *Eur. J. Oper. Res.* **197**(1) 415–421.
Wang, X., L. G. Debo, a. Scheller-Wolf, S. F. Smith. 2010. Design and analysis of diagnostic service centers. *Management Sci.* **56**(11) 1873–1890.
Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. MIT Press, Cambridge, Massachusetts.
Wooldridge, J. M. 2012. *Introductory Econometrics: A Modern Approach*, 5th ed. Mason, OH, South-Western Cengage Learning.
Zeckhauser, R. J. 1970. Medical insurance: A case study of the tradeoff between risk spreading and appropriate incentives. *J. Econom. Theory* **2** 10–26.

Figure 1. Standard Patient Flow in the Emergency Department



[1] for patients who were admitted to the hospital
[2] for patients who were discharged to home or an outside facility

Table 1. Summary definition of variables

Variable	Description	Level of Analysis
Main dependent variable		
Length of stay	Logged number of minutes for which patients stayed in ED.	Patient
Independent and control variables		
ESI level	4 indicators for patient's ESI level (from highest to lowest: 2, 3, 4, 5).[§]	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 dataset.	Physician-shift
Fast others	Indicator for presence of above-average productivity physician on shift (= 1 for present, = 0 for absent).	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	3 indicators for type of shift during which patient arrived (AM, PM, overnight).	Patient
Months since March 2007	Indicator for what number month this is in this dataset.	Patient
Month	12 indicators for month of shift.	Patient
Day of week	7 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 pre-implementation, = 0 for post-implementation).	Physician-shift
Interaction	PAS × Main ED.	Physician-shift
Additional dependent variables		
Discharge rate	Number of patients discharged per hour by a given physician in a given 2-hour period of the shift (e.g., penultimate 2 hours, final 2 hours).	Physician-shift (2-hour period)
ED wait time	Logged number of minutes elapsed between patient arrival to ED and MD start.	Patient
ED care delivery time	Logged number of minutes elapsed between MD start and bed request (for patients admitted to hospital) or discharge from ED (for patients discharged to home or an outside facility).	Patient
ED boarding time	Logged number of minutes elapsed between bed request and discharge from ED (if admitted to hospital).	Patient
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 upon 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
Shift duration	Number of hours for which physician worked in ED during this shift.	Physician-shift

[§] Although the Emergency Severity Index (ESI) uses five categories, we have four indicators for patient ESI level because we exclude patients of ESI level 1 from our analysis.

Table 2. Summary statistics of variables included in models

Variable	Overall				Pre-PAS				Post-PAS			
	Mean	SD	Min.	Max.	Mean	SD	Min.	Max.	Mean	SD	Min.	Max.
1. Length of stay (minutes)	198.6	199.1	1	2879	202.9	191.5	1	2879	195.4	203.3	1	2878
2. Discharge rate	2.4	1.3	0.5	9.5	2.4	1.3	0.5	8.5	2.4	1.3	0.5	9.5
3. ED wait time (minutes)	41.4	37.2	1	1452	47.4	43.4	1	1441	37.4	31.9	1	1452
4. ED care delivery time (minutes)	159.4	201.7	0	2867	175.5	200.4	0	2867	171.6	209.9	0	2864
5. ED boarding time (minutes)	218.4	313.5	0	2685	301.1	378.2	0	2631	152.2	224.0	0	2456
6. Age (years)	37.8	24.4	0	112	38.3	24.1	0	111	37.4	24.6	0	112
7. MDs on shift	4.6	1.2	1	9	4.4	1.1	1	7	4.7	1.3	1	9
8. Current waiting count	2.4	1.8	0	21	2.5	2.0	0	21	2.3	1.7	0	17
9. Current patient count	5.5	2.9	1	27	5.6	3.0	1	27	5.5	2.8	1	23
10. Shift number	254.6	162.8	1	678	121.8	74.7	1	308	347.4	144.9	1	678
11. Shift duration (hours)	9.6	1.4	2	27.6	9.8	1.4	2	25.7	9.5	1.3	2	27.6

Variable	1	2	3	4	5	6	7	8	9	10	11
1. Length of stay (minutes)	1										
2. Discharge rate	-0.14*	1									
3. ED wait time (minutes)	0.05*	0.22*	1								
4. ED care delivery time (minutes)	0.98*	-0.22*	-0.13*	1							
5. ED boarding time (minutes)	0.42*	-0.24*	0.07*	0.42*	1						
6. Age (years)	0.28*	-0.17*	-0.12*	0.30*	0.06*	1					
7. MDs on shift	-0.04*	0.14*	0.08*	-0.05*	-0.02*	-0.03*	1				
8. Current waiting count	-0.10*	0.52*	0.50*	-0.19*	0.06*	-0.18*	0.13*	1			
9. Current patient count	0.02*	0.45*	0.33*	-0.04*	0.01	-0.07*	-0.08*	0.61*	1		
10. Shift number	-0.06*	0.08*	-0.07*	-0.04*	-0.16*	-0.05*	0.21*	0.01*	0.00	1	
11. Shift duration (hours)	-0.10*	0.19*	0.11*	-0.11*	-0.00	-0.06*	0.12*	0.17*	0.10*	-0.04*	1

* $p < 0.05$

Table 2 (continued). Summary statistics of variables included in models

Variable	Overall	Pre-PAS	Post-PAS
ESI level 2	8.01	5.27	9.73
ESI level 3	51.32	51.50	51.33
ESI level 4	39.25	41.78	37.59
ESI level 5	1.42	1.45	1.36
Other fast MDs on shift	86.85	86.93	86.87
ESI level 1 patient present	9.26	9.09	9.38
Trauma patient present	19.44	7.43	28.12
AM shift	36.47	36.51	36.44
PM shift	47.35	47.46	47.28
Overnight shift	16.18	16.03	16.28
Main ED	67.88	66.64	68.79
PAS implemented	59.04	0	100
2007	22.92	57.29	--
2008[§]	28.07	42.71	15.03
2009[§]	30.68	--	53.19
2010	18.33	--	31.78
January	7.40	5.96	8.69
February	7.36	6.39	8.33
March[§]	10.37	12.30	9.44
April[§]	9.83	11.55	9.04
May[§]	10.46	12.12	9.73
June[§]	9.68	11.51	8.79
July[§]	10.01	11.87	9.13
August	7.26	5.83	4.53
September	7.11	5.65	8.41
October	7.06	5.63	8.33
November	6.74	5.54	7.84
December	6.73	5.64	7.75
Sunday	15.00	15.07	14.87
Monday	15.05	14.81	15.26
Tuesday	14.03	14.08	14.04
Wednesday	13.61	13.91	13.46
Thursday	13.76	13.89	13.71
Friday	13.81	13.80	13.79
Saturday	14.75	14.43	14.87
Lab ordered	47.33	45.48	48.53
X-ray ordered	35.11	34.53	35.47
Admitted to hospital	9.77	10.24	9.46

Note: N = 231,081. Excludes observations earlier than March 1, 2007 and after July 31, 2010.

[§] Because all observations earlier than March 1, 2007 and after July 31, 2010 have been excluded, it is not surprising that a larger percentage of patients in our dataset presented to the ED in the months between March and July (inclusive) and in the years of 2008 and 2009, respectively. When these summary statistics are produced with the inclusion of observations all from January 1, 2007 to December 31, 2010, we obtain an approximately uniform distribution of patients across all months of the year.

Table 3. Average length of stay in minutes by patient ESI level

ESI level	Mean	SD	Frequency
2 (most severe)	402.84	359.53	18499
3	279.44	262.67	118602
4	108.33	97.67	90688
5 (least severe)	83.11	90.55	3292

N = 231,081.

Table 4. Fixed effects models at patient level

Variables	(1) Logged ED Length of Stay	(2) Logged ED Wait Time	(3) Logged ED Care Delivery Time	(4) Logged ED Boarding Time
ESI level 3	-0.274***	0.415***	-0.398***	0.119***
	(0.0129)	(0.0133)	(0.0158)	(0.0164)
ESI level 4	-0.736***	0.696***	-1.207***	0.0131
	(0.0137)	(0.0242)	(0.0200)	(0.0714)
ESI level 5	-0.967***	0.617***	-1.578***	0.305
	(0.0182)	(0.0289)	(0.0251)	(0.742)
Age	0.00551***	-0.00263***	0.00772***	0.00463***
	(0.000157)	(0.000184)	(0.000235)	(0.000464)
MDs on shift	0.00260	0.0209*	-0.00582*	0.0123
	(0.00337)	(0.00810)	(0.00261)	(0.0134)
Current waiting count	0.1000***	0.190***	0.00148	0.0590***
	(0.00462)	(0.00573)	(0.00181)	(0.00898)
Current patient count	0.00362*	0.0199***	0.00131	-0.00201
	(0.00176)	(0.00480)	(0.00171)	(0.00347)
Shift number	-0.000306*	-0.0000063	-0.000553*	-0.000915*
	(0.000144)	(0.000358)	(0.000247)	(0.000338)
Fast others	-0.00231	-0.0354***	0.0145	-0.0372
	(0.00605)	(0.00750)	(0.00769)	(0.0292)
ESI level 1 patient present	0.0245***	0.0642***	0.0145*	0.0689**
	(0.00454)	(0.00774)	(0.00559)	(0.0227)
Trauma patient present	0.0280***	0.0630***	0.0147*	-0.00235
	(0.00367)	(0.00594)	(0.00594)	(0.0247)
PM shift	-0.0143	0.0698***	-0.0613***	-0.0261
	(0.00909)	(0.0160)	(0.00711)	(0.0249)
Overnight shift	-0.0880***	-0.156***	-0.0701***	0.173***
	(0.0131)	(0.0287)	(0.0135)	(0.0349)
Months since March 2007	0.00347	0.00239	0.00567	-0.0123*
	(0.00192)	(0.00478)	(0.00337)	(0.00474)
Main ED	0.561***	0.380***	0.643***	0.438*
	(0.0299)	(0.0338)	(0.0318)	(0.181)
PAS	-0.0253	-0.174***	0.0870**	0.0619
	(0.0150)	(0.0400)	(0.0259)	(0.150)
PAS x Main ED	-0.100***	-0.0853**	-0.174***	-0.236
	(0.0137)	(0.0260)	(0.0214)	(0.152)
Constant	4.677***	2.270***	4.556***	4.383***
	(0.0406)	(0.0516)	(0.0375)	(0.218)
Observations	217,810	219,305	217,161	21,425
Number of ED physicians	40	40	40	40
Adjusted R^2	0.374	0.282	0.440	0.114

*** p <0.001, ** p <0.01, * p <0.05

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

Table 5. Fixed effects models at physician-shift levels

Variables	(1) Discharge Rate	(2) Shift Duration
Percent of ESI level 3 patients	0.00238*** (0.000505)	0.0126*** (0.00162)
Percent of ESI level 4 patients	0.00667*** (0.000818)	0.0311*** (0.00240)
Percent of ESI level 5 patients	0.00587* (0.00257)	0.0315*** (0.00530)
Average age of patients	0.00103** (0.000373)	-0.0203*** (0.00312)
MDs on shift	-0.0398*** (0.00618)	-0.332*** (0.0243)
Average waiting count	0.0172 (0.0113)	-0.519*** (0.0360)
Average patient count	0.163*** (0.00746)	0.650*** (0.0256)
Shift number	0.000291 (0.000204)	0.000189 (0.000803)
Fast others	0.00331 (0.0126)	-0.0710 (0.0465)
Percent of time ESI level 1 patient present	0.00994 (0.0164)	0.0687 (0.0516)
Percent of time trauma patient present	0.0105 (0.0127)	-0.0232 (0.0437)
PM shift	0.219*** (0.0155)	-0.130* (0.0575)
Overnight shift	0.0625* (0.0233)	-1.864*** (0.138)
Months since March 2007	-0.00360 (0.00298)	0.0164 (0.0113)
Main ED	--	1.035*** (0.171)
PAS	0.116*** (0.0263)	-0.465*** (0.0945)
PAS x Main ED	--	-0.103 (0.0901)
Discharged in penultimate 2 hours of shift	0.00403 (0.0191)	--
PAS x discharged in penultimate 2 hours of shift	0.0467* (0.0203)	--
Constant	0.389*** (0.0828)	7.200*** (0.327)
Observations	21,582	14,155
Number of ED physicians	40	40
Adjusted R^2	0.161	0.296

*** p <0.001, ** p <0.01, * p <0.05

Note: Both regressions include month controls, day of week controls, and fixed effects for each physician. Model (1) is estimated at the physician-shift 2-hour period level, and model (2) is estimated at the physician-shift 2-hour period level. Discharge rate in model (1) reflects the number of patients discharged per hour by a given physician in a given 2-hour period of the shift. Shift duration in model (2) is expressed in hours. Standard errors (in parentheses) are heteroskedasticity robust and clustered by physician.

Table 6. 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) Died in ED
ESI level 2	--	--	--	-4.525***
				(0.207)
ESI level 3	-0.678***	-0.519***	-0.923***	-6.766***
	(0.0328)	(0.0329)	(0.0282)	(0.319)
ESI level 4	-2.546***	-0.808***	-2.783***	-8.084***
	(0.0430)	(0.0475)	(0.0726)	(0.953)
ESI level 5	-3.284***	-2.351***	-4.784***	--
	(0.0728)	(0.117)	(0.992)	
Age	0.0177***	0.0222***	0.0348***	0.0293***
	(0.000652)	(0.000949)	(0.000662)	(0.00426)
MDs on shift	-0.0316	-0.0251*	-0.00517	-0.0492
	(0.0169)	(0.0111)	(0.0124)	(0.111)
Current waiting count	0.0137	0.00696	-0.0182	-0.0837
	(0.00744)	(0.00572)	(0.00971)	(0.0777)
Current patient count	-0.0161***	0.00206	-0.00163	0.00892
	(0.00461)	(0.00388)	(0.00421)	(0.0306)
Shift number	-0.000417	-0.000471	-0.000368	0.000178
	(0.000304)	(0.000257)	(0.000258)	(0.000733)
Fast others	0.0815***	0.0105	0.0429	-0.172
	(0.0245)	(0.0221)	(0.0280)	(0.210)
ESI level 1 patient present	0.0320	-0.00363	-0.0382	0.714**
	(0.0281)	(0.0161)	(0.0304)	(0.235)
Trauma patient present	-0.0180	0.0103	0.00706	-0.174
	(0.0189)	(0.0158)	(0.0251)	(0.169)
PM shift	-0.103	0.0287	-0.00139	0.218
	(0.0546)	(0.0272)	(0.0320)	(0.157)
Overnight shift	-0.176***	-0.0251	0.0113	0.498
	(0.0526)	(0.0326)	(0.0417)	(0.281)
Months since March 2007	0.000203	0.00603	-0.0145***	0.0165
	(0.00460)	(0.00323)	(0.00396)	(0.0162)
Main ED	1.443***	-0.104*	1.529***	--
	(0.114)	(0.0504)	(0.102)	
PAS	0.194***	0.0325	0.252	-0.792*
	(0.0461)	(0.0386)	(0.156)	(0.366)
PAS × Main ED	-0.0802	-0.0271	-0.169	--
	(0.0461)	(0.0385)	(0.145)	
Constant	-0.279	-0.668***	-3.771***	-2.636***
	(0.153)	(0.0640)	(0.121)	(0.570)
Observations	196,134	196,134	196,134	135,280
Pseudo R^2	0.331	0.0682	0.238	0.564

*** p <0.001, ** p <0.01, * p <0.05

Note: All regressions are estimated at the patient level and include month controls and day of week controls. Model (4) 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.