Doctors Under Load: An Empirical Study of State-Dependent Service Times in Emergency Care

Robert J. Batt
Wisconsin School of Business, University of Wisconsin-Madison, Madison, WI 53706, rbatt@bus.wisc.edu
Christian Terwiesch
The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, terwiesch@wharton.upenn.edu

This paper studies mechanisms which cause service times in a complex service system to shift in response to changes in the system occupancy level. Using operational data from over 140,000 patient visits at a hospital emergency department, we show that as crowding increases, the amount of time a patient spends in the treatment room first increases then decreases. We show that this inverted U-shaped response is driven by multiple mechanisms working simultaneously. Increasing workload leads to queuing delays for individual resources, such as nurses, which retards patient treatment. However, using a novel dataset of 5.4 million real-time nurse location tracking observations, we find that nurses attempt to counteract this slowdown by reducing the amount of time spent with patients. We also propose and provide evidence of *early task initiation* being used to speed up patient service times. This occurs when an upstream server (triage nurse) starts tasks (diagnostic tests) early in a patient's visit that normally would be started later by a downstream server (physician). Early task initiation, however, is potentially costly as it can lead to an increase in total tasks due to the upstream server initiating unneeded tasks. Lastly, we use simulation to show that ignoring load-dependent service times leads to modeling errors that could cause hospitals to overinvest in human and physical resources.

Key words: Healthcare operations; empirical; emergency department; tandem queue
History: Under Revision for Management Science

1. Introduction

The Operations Management community has long been concerned with how crowding affects the performance of queuing systems. Basic queuing theory shows that crowding and high utilization of queues lead to exponentially increasing wait times. Since long waits are generally undesirable, it seems reasonable that, when possible, workers in human-paced service systems would attempt to accelerate the system, a phenomenon we call *Speedup*. Indeed, this has been shown to be true both in the lab and in practice (Schultz et al. 1998, Kc and Terwiesch 2009, Chan et al. 2011). These papers show that workers in settings as varied as data-entry and hospital intensive care units accelerate service under high workload conditions.

In contrast, in domains such as transportation and telecommunications, high workload conditions are well known to lead to service time increases or *Slowdown* (Chen et al. 2001, Gerla and Kleinrock 1980). A hallmark of Slowdown-prone systems is that service involves shared or multitasking resources wherein a server serves several customers either simultaneously or in rapid succession. For example, a highway lane is a shared resource for all the cars traveling in it, and similarly, each node in a telecom network is a shared resource for many users (Gerla and Kleinrock 1980).

We bring these viewpoints together by empirically analyzing a service system where both Speedup and Slowdown effects are present: a hospital emergency department (ED). The ED provides an excellent study environment for two reasons. First, the ED is an environment where the servers (doctors, nurses, technicians, etc.) have a great deal of discretion over the content and timing of the encounter, leading to the potential for service Speedup. Second, the ED is an environment where many resources are shared across patients, leading to the potential for service Slowdown.

Many papers in the medical literature have shown the negative impacts of ED crowding on such measures as timing of antibiotic delivery for pneumonia patients, pain medication for patients with severe pain, and nebulizer treatment for patients with asthma (Pines et al. 2006, Fee et al. 2007, Pines and Hollander 2008, Pines et al. 2010). Crowding has also been associated with reduced patient satisfaction (Pines et al. 2008). Results on the impact of crowding on length of stay have been mixed. For example, Pines et al. (2010) reports a positive relationship between crowding and length of stay while Lucas et al. (2009) finds no significant relationship. McCarthy et al. (2009) reports that crowding drives up wait times but has no effect on service times, a result that agrees with traditional queuing theory.

In this paper, we identify and test for several mechanisms which may cause the service rate to change in response to the busyness of the system. These include shared resource queuing delays, rushing, task reduction, and early task initiation. The first is a Slowdown mechanism and the other three are Speedup mechanisms. Understanding such effects is critical to managing service systems because they impact the customer experience directly and, as we show, they potentially have an impact on capacity requirements. From an academic perspective, examining Speedup and Slowdown mechanisms opens the black box of a worker-driven queuing system and shows that the behavior of those working in the system can be impacted by the system load in a multitude of ways, with different mechanisms working simultaneously.

We conduct a detailed econometric analysis of service duration and service content for more than three years of emergency department visits at a major U.S. hospital. From hospital electronic medical record data we observe patient-level data such as age, gender, race, and diagnostic testing orders. We also incorporate a novel dataset from a nurse location-tracking system that allows us to observe when nurses enter and exit patient rooms. To the best of our knowledge, this is the first

paper to incorporate micro level data about the time care providers spend with patients. Duration analysis models are used to estimate the effects of system busyness on various duration measures. Count-based models are used to estimate the effects of system busyness on the number of diagnostic test orders. Lastly, we use discrete event simulation to determine if the identified state-dependencies have a managerially meaningful impact on the system. This research design allows us to make the following four contributions:

1. We provide evidence of a complex service system with multitasking resources exhibiting non-monotone service times with respect to system busyness, first increasing then decreasing. We find that the mean in-service time for an abdominal pain patient first increases from an average of 5.1 hours to 5.6 hours and then drops back down to 5.4 hours.

2. We show that the queuing delays created by increasing workload on shared resources causes Slowdown in the completion time of tasks such as medication delivery and lab specimen collection. Factors such as the number of medications and lab tests ordered per hour, and the number of patients being treated all lead to delays in task completion. For example, a one unit increase in the number of medications ordered per hour corresponds to a contemporaneous 2.9% increase in medication delivery time. Further, we show that one type of shared resource, nurses, exhibit rushing behavior but that it is insufficient to overcome the workload-induced delays. We find that mean nurse-patient interaction time drops 0.2 minutes per patient hour (approximately 4%) with a one unit increase in the waiting room census.

3. We show that as the ED gets busier, triage nurses order more diagnostic tests and doctors order correspondingly fewer tests. Shifting a diagnostic test earlier in the patient visit (early task initiation) reduces service time by approximately 10 minutes on average, but can also to lead to an increase in total diagnostic testing, which erodes this time savings.

4. We show that models which ignore the state-dependent nature of service times overestimate system performance metrics such as wait time and length of stay by 10% to 20%.

These findings offer operational insights for managers as well. For example, we show that implementing early task initiation by increasing the number of tests ordered at triage is an effective way to reduce service time, particularly when the system is under high load. This suggests that care providers should consider incorporating state-dependencies into ED care protocols. Further, while the results we present are specific to the ED, our theoretical foundations of state-dependent service times as well as our methodologies for disentangling the effects are broadly applicable to other worker-driven service systems. Our findings show that understanding the micro-level mechanisms behind state-dependent service times is important for properly modeling and managing service systems.

2. Clinical Setting

Our study is based on data from a large, urban, teaching hospital with an average of 4,700 ED visits per month. The study ED has 33 treatment rooms and 7 hallway beds for a theoretical maximum treatment capacity of 40 beds. However, the actual treatment capacity at any given moment can fluctuate for various reasons (e.g. room out of service, extra hallway bed). The hospital also operates an express lane or FastTrack (FT) for low acuity patients. The FT is generally open from 8am to 8pm on weekdays, and from 9am to 6pm on weekends. The FT shares the waiting room with the main ED, but other than that, it operates largely autonomously from the rest of the ED. It utilizes seven dedicated beds and is usually staffed by a dedicated group of Certified Registered Nurse Practitioners rather than Medical Doctors[1] .

In our analysis, we focus solely on patients that are classified as "walk-ins" or "self" arrivals, as opposed to ambulance, police, or helicopter arrivals. This is because the walk-ins go through a more standardized process of triage, waiting, and treatment, as described below. In contrast, ambulance arrivals tend to jump the queue for bed placement, regardless of severity, and often do not go through the triage process or wait in the waiting room. More than 70% of ED arrivals are walk-ins. Note, however, that the non-walkin patients are included in the relevant census measures. Similarly, for clarity, we focus on patients treated in the main ED rather than in the FT. Again, the FT patients are included in all relevant census measures

The study hospital operates in a manner similar to many hospitals across the United States. Each patient visit can be broken down into three main phases: Waiting Room, In-Service, and Boarding. The Waiting Room Phase begins upon arrival when patients are checked in and an electronic patient record is initiated for that visit. Only basic information (name, age, complaint) is collected at check-in. Shortly thereafter, the patient is seen by a triage nurse who assesses the patient, measures vital signs, and records the chief complaint. The triage nurse also assigns a triage level which indicates acuity. The hospital uses a five-level Emergency Severity Index triage scale with 1 being most severe and 5 being least severe (Gilboy et al. 2011). The triage nurse also has the option of ordering pathology lab tests (e.g., urinalysis, blood test) and certain types of radiology imaging scans (e.g., x-rays)[2]. These tests are carried out by other nurses or technicians after the triage process is complete.

After triage, all patients wait in a common waiting room to be taken to a treatment room. If tests were ordered by the triage nurse, the patient will be temporarily removed from the waiting room

[1] We interchangeably use the term ED to refer to the entire Emergency Department inclusive of the FastTrack or to just the main emergency department treatment area exclusive of the FastTrack. The use is generally clear from the context, but we use the term "main ED" to clarify and indicate the primary ED treatment space when necessary.

[2] These triage-ordered tests are commonly referred to as Advance Triage Protocols.

to execute the test (e.g. have blood drawn). Patients are called for service when a treatment bed is available. If only the ED is open, patients are generally (but not strictly) called for service in first-come-first-served (FCFS) order by triage level. If the FT is open, then the FT will generally serve triage level 4 and 5 patients in FCFS order by triage level and the ED will serve patients of triage levels 1 through 3 in FCFS order by triage level. These routing procedures are flexible, however. The mean and median wait times for ED patients are are 1.6 hours and 0.84 hours, respectively. The mean and median wait times for FT patients are 1.1 hours and 0.9 hours, respectively.

Patients served by the main ED are eventually assigned to a treatment room by the charge nurse.[3] This transition marks the end of the Waiting Room Phase and the beginning of the In-Service Phase. Soon after being moved to a treatment room, a physician meets with and examines the patient.[4] Each physician is responsible for a fixed block of treatment rooms and treats whichever patients are assigned to those rooms by the charge nurse. At this point, the physician generates a mental list of possible diagnoses, called a differential diagnosis, and decides the trajectory of the diagnosis and treatment process. Frequently, orders for diagnostic tests (e.g. pathology lab tests, radiology imaging), medications, or both are made at this point. All lab test, radiology scan, and medication orders are recorded electronically in the patient tracking system. Orders are frequently conveyed orally to the nurses as well.

Lab specimens are drawn by the nurse and most are sent to the hospital's central pathology lab by pneumatic tube for processing. A small subset of pathology tests are performed locally in the ED by the nurse. Similarly, the nurse is responsible for delivering medications to the patient. When the nurse finishes either of these tasks, the order is closed out and timestamped in the electronic patient record. Orders for radiology scans trigger a patient-transport request. Transporters work in a first-come-first-served manner through the request queue to transport patients to the appropriate imaging equipment and then back to the treatment room.

This process of physician visit, order generation, and order completion may happen once or several times for each patient. Eventually, the physician decides that either the patient can be discharged from the ED or the patient needs to be admitted to the hospital. If the patient is to be admitted, a bed request is entered in the inpatient bed management system. For the admitted patient, this marks the end of the In-Service Phase and the beginning of the Boarding Phase. The patient waits in the ED for an available inpatient bed and is considered a "boarder." The Boarding Phase can be quite long with a mean of 3.6 hours. During this time, the patient continues to occupy a treatment

[3] The treatment location is sometimes a hallway bed rather than a room, but we use the word "room" for ease of exposition.

[4] Because the study hospital is a teaching hospital, a medical student or a resident physician may also be involved in the care of the patient.

room and requires some attention from the nursing staff, but the physician is effectively done with the patient. The number of boarding patients in the ED ranges from zero to 20 with a mean of six. For patients that are discharged, the In-Service Phase ends when the patient leaves the ED (there is no Boarding Phase for discharged patients). Mean In-Service time for admitted and discharged patients is 3.6 hours and 3.8 hours respectively.

Note that in-service time is a measure of the time a patient is occupying a treatment space, not a measure of the amount of value-added time performed for the patient. Additionally, the activities included in the In-Service Phase are slightly different for admitted and discharged patients. For discharged patients, the discharge process (e.g., final instructions from the nurse) is included in the In-Service time. For admitted patients, there is an admitting process rather than a discharge process, but this generally occurs after the bed request has been submitted and thus is included in the Boarding Phase time rather than the In-Service Phase time. We control for this difference by including a dummy variable for admitted patients in all relevant analyses.

3. Data Description

This study uses data from two sources at the hospital. The main dataset is from the ED electronic medical record system and includes all ED patient visits over the period of January, 2009 through December, 2011, approximately 140,000 patient visits. The data include information for each patient visit such as patient demographics, chief complaint, attending physician, and timestamps of all major events and physician orders. Table 1 provides descriptive statistics of the patient population.

Table 1 Summary Statistics of ED Patients

Variable	All Patients Mean	Abdominal Pain Patients Mean
Age	39.8 (0.05)	38.0 (0.14)
Female	60% (0.001)	70% (0.004)
Triage 2	19.1% (0.001)	16.7% (0.003)
Triage 3	45.3% (0.001)	81.6% (0.003)
Race: Black	60.0% (0.001)	59.1% (0.004)
Race: White	23.8% (0.001)	24.5% (0.004)
Diagnostics Ordered	4.0 (0.01)	6.7 (0.04)
Wait Time (hr.)	1.5 (.01)	1.8 (0.02)
Service Time (hr.)	3.1 (0.01)	4.9 (0.03)
Boarding Time† (hr.)	3.0 (0.02)	2.8 (0.05)
Total Time (hr.)	5.2 (0.01)	7.4 (0.03)
N	141,616	13,802

Standard error in parentheses

†Boarding Time conditional on being admitted

Over 200 unique chief complaints occur in the dataset, ranging from chest pain to intoxication. Because the care process for different chief complaints can be quite varied, comparing system performance metrics across all chief complaints potentially produces results that are "on average" correct, but are not representative of any real patient. Therefore, for the sake of clarity, in this study we focus on patients with the chief complaint of abdominal pain unless otherwise indicated. Abdominal pain is the most common chief complaint among non-FastTrack patients, occurring in 13% of the patient visits (or just under 10% of total ED and FT visits). Abdominal pain is also an attractive focal complaint for this study because it is a clinically vague complaint and thus there can be a wide range of diagnostic paths ordered by the physician. This is potentially favorable for finding care that changes in response to system busyness. Additionally, because 98% of abdominal pain patients are classified as triage level ESI 2 or ESI 3, we drop the other triage levels from most analyses.

One limitation of the dataset is that we do not observe staffing levels. Controlling for staffing is important as we study the effects of workload on system service times since the impact of a given level or workload may be different as it is served by varying numbers of care providers. The study hospital follows a fairly rigid staffing plan such that the staffing on a given shift of a given day of the week if quite consistent. For example, on Mondays there are usually two attending physicians on duty from 7:00am to 3:00pm and three on duty from 3:00pm to 11:00pm. The hospital does very little real-time demand based staffing. Therefore, we use time-related variables, such as day-of-week and workshift or hour-of-day, to control for staffing patterns whenever possible.

With an increasing usage of electronic medical records and patient flow systems in hospitals, such patient level data is increasingly used in the academic literature (e.g., Pines et al. 2006, Kc and Terwiesch 2009, Chan et al. 2011). However, such patient level data provides little information about the human resource consumption of these patients. To the best of our knowledge, ours is the first study that accurately measures the amount of time care-providers actually spend in the room with the patient. To do this, we draw on data from a second source, the nurse tracking and communication system. This system, installed in early 2012, uses a network of 146 infrared sensors located in the walls and ceiling throughout the ED to track the physical location of nurses as they go about their work. Each nurse wears a small infrared transmitter tag that emits an identification signal every few seconds which is picked up by the sensors in the ED. This tracking capability is part of the telephone/communication system and its main purpose is to allow calls and messages to be routed directly to a nurse's location. For our purposes, the system allows us to observe how much time nurses spend in direct interaction with the patient in the treatment room. Due to data archiving limitations of the system, data prior to May, 2013 is not available. Our dataset includes 5.4 million location observations from May, 2013 through September, 2013.

We also obtain the matching patient-level data from the electronic medical record system, similar to the main dataset described above, for the May, 2013 through September, 2013 time period. Putting these two data sources together allows us to calculate such measures as the the number of nurse-patient interactions, the mean interaction duration, and the total nurse-patient interaction time for each patient visit which occurs in a treatment room. (We cannot reliably compute these metrics for patients in hallway beds because the tracking system cannot differentiate between a nurse interacting with a hallway patient and a nurse standing in the hallway for some other reason. The vast majority of patient visits take place in treatment rooms, so this is a minor loss.)

Table 2 Summary statistics of nurse-patient interactions

Variable	ESI 2 Mean	ESI 3 Mean
Interactions per Patient Visit	15.3 (0.52)	9.5 (0.22)
Duration of an Interaction (minutes)	3.1 (0.07)	2.9 (0.05)
Total Interaction Time per Patient Visit (minutes)	45.8 (1.80)	26.6 (0.76)
Interactions per Patient Hour	2.2 (0.07)	1.9 (0.03)
Interaction Time per Patient Hour (minutes)	6.9 (0.44)	5.4 (0.12)
# of Unique Nurses Interacting with Patient	4.9 (0.14)	3.3 (0.06)
N	406	1,375

Standard error in parentheses

Table 2 provides summary statistics of these measures for abdominal pain patients of triage levels ESI 2 and ESI 3. On average, nurses spend 45.8 minutes and 26.6 minutes with ESI 2 and ESI 3 patients, respectively. We normalize this measure by dividing by the amount of time the patient is in a treatment room and find that ESI 2 and ESI 3 patients received 6.9 minutes and 5.4 minutes of direct nurse interaction time per hour of patient time spent in a treatment room.

4. Analysis of Patient Visit Segment Durations

We are interested in examining the mechanisms of state-dependent services times, but we must first determine if such a phenomenon exists. We begin with the assumption from classical queuing theory that the service time distribution is not affected by the system state (Wolff 1989). Under such an assumption, if we view the entire ED process as a "black box," by Little's Law we expect total length of stay (LOS) in the ED to be linearly increasing with the number of people in the ED (Little 1961). If we look "inside the black box" and consider the Waiting Room Phase and the In-Service Phase of the ED to be akin to the waiting and service portions of a simple multi-server queue, then again by Little's Law we expect waiting room time to be linearly increasing in the number of people in the waiting room and by assumption we expect in-service time to be unchanging in either waiting room or in-service census.

4.1. Model Formulation

We are interested in the change in a duration in response to a measure of system load (e.g. mean LOS with respect to total ED census). To do this we use an accelerated-failure-time (AFT) model, which is a form of parametric duration regression (Greene 2012, Sec. 19.4.3). The AFT model relates the log of a duration to a vector of covariates and a random error term ϵ through a linear equation. This model takes the general form

$$ln(Duration_i) = \alpha + \beta_1 Census_i + \beta_2 Census_i^2 + \mathbf{W}_i\boldsymbol{\theta} + \mathbf{Z}_i\boldsymbol{\phi} + \epsilon_i \quad (1)$$

where the index i denotes a patient visit to the ED. We estimate Equation 1 for four pairs of *Duration* and *Census* variables:

1. Total LOS vs. Total ED Census
2. Wait Time vs. Waiting Room Census
3. In-Service Time vs. In-Service Census
4. In-Service Time vs. Waiting Room Census

We choose these particular variable pairs for a few reasons. The pairs in the first three models represent the logical pairing that would be used in a typical Little's Law analysis. Also, these represent the three points of view described above in Section 4. The fourth pair may at first seem surprising, however we choose to test in-service time versus the waiting room census because our observations of the ED and our conversations with care providers have shown that the waiting room census is the main crowding metric on which the care providers focus. Therefore we suspect this metric drives a behavioral response in the care providers. We include both the linear and quadratic forms of the census variable to allow for nonmonotonic response to census. $\mathbf{W}_i$ is a vector of patient-visit specific covariates including age, gender, race, triage level, admitted status, and physician. $\boldsymbol{Z}_i$ is a vector of time-related control variables including year, month, hour of day, weekend indicator, and the interaction of hour of day and weekend.

In AFT models, the distribution assumption of the error term implies specific characteristics of the underlying hazard function of the data. We assume ϵ is logistically distributed which allows for a hazard function that can be either monotonically decreasing or nonmonotonic, first increasing and then decreasing. We choose this distribution because this form resembles the hazard function form of the data and because it provides a better model fit, based on the Bayesian Information Criterion, than does the more common normal assumption (e.g. Kc and Terwiesch 2009, Tan and Netessine 2012).

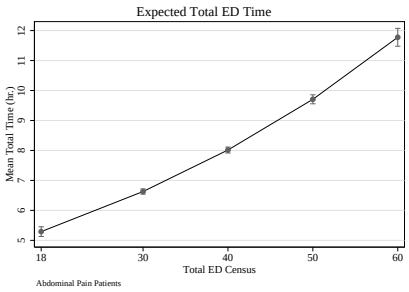
Table 3 Impact of census on elements of patient visit duration

	(1)	(2)	(3)	(4)
Duration Variable	Total Time	Wait Time	In-Service Time	In-Service Time
Census Variable	Total ED (upon arrival)	Waiting Room (upon arrival)	In-Service (upon start of service)	Waiting Room (upon start of service)
Census	0.0184***	0.2814***	0.0175***	0.0135***
	(0.0022)	(0.0039)	(0.0065)	(0.0030)
Census^2	0.0000	-0.0063***	-0.0003	-0.0005***
	(0.0000)	(0.0001)	(0.0002)	(0.0001)
N	12,457	12,457	12,261	12,457

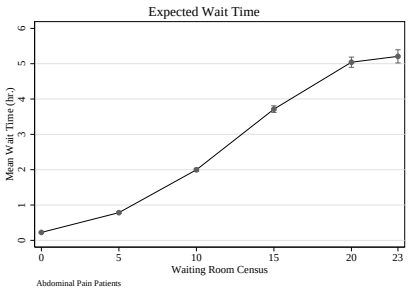
Robust standard errors in parentheses
Controls not shown: age, gender, race, triage level, admit, physician, year, month, hour×weekend
Accelerated failure time model with loglogistic distribution
* $p<0.10$, ** $p<0.05$, *** $p<0.01$

Figure 1 Duration elements as a function of census

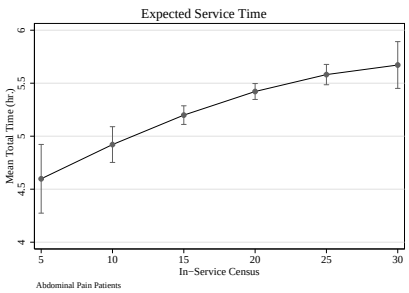
(a) Total ED Time vs. Total ED Census



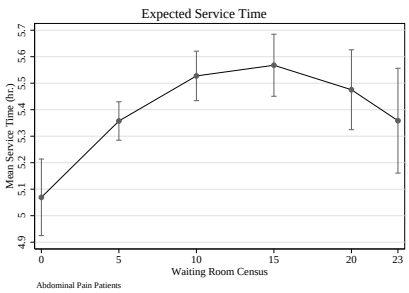
(b) Waiting Room Time vs. Waiting Room Census



(c) In-Service Time vs. In-Service Census



(d) In-Service Time vs. Waiting Room Census



Note: Error bars indicate 95% confidence interval of the mean.

4.2. Results

The estimation results of the AFT duration models are shown in Table 3. The nonlinear nature of the model and the inclusion of the quadratic census term make direct interpretation of the estimated coefficients difficult. Plots of predicted values clarify the results.

Figure 1 displays plots of expected durations predicted by the models. Figure 1a shows that, as anticipated, the expected total ED length of stay increases approximately linearly from 5.3 hours to 11.8 hours as total ED census ranges from its 5th percentile to 95th percentile value. Similarly, Figure 1b shows that expected wait time increases from 0.25 hours to over five hours as the waiting room census ranges from its 5th percentile to 95th percentile value. However, the trend is nonlinear, suggesting that the traditional queuing assumptions do not hold in the study ED.

Figure 1c shows that the classic assumption of fixed services times does not hold in the ED. Rather, service times increase from about 4.6 hours to 5.6 hours as the in-service census varies over its typical range. While this increase does not conform to traditional queuing assumptions, it certainly conforms to the common sense hypothesis that service slows down when the system is busier.

Figure 1d is a surprising contrast to the other plots of Figure 1. However, the inverted-U shape helps explain Figure 1b and provides the motivation for the rest of this study. Figure 1d shows that rather than being either invariable or monotone increasing, mean in-service time increases with waiting room census up to about the 75th percentile of census (15 people) and then decreases. This pattern explains the sigmoid shape of Figure 1b. As the census shifts from low to moderate, the in-service times increase causing a convex increase in wait times. As the census increases from moderate to high, in-service times decrease, leading to the inflection in the wait time response.

While it is possible that a single phenomenon is causing the nonmonotone, concave shape of Figure 1d we are not aware of any prior literature suggesting such a mechanism. Rather, we hypothesize that multiple mechanisms are at work and it is the combined effect of these mechanisms that leads to the inverted-U response of in-service time to census. The remainder of this paper examines possible mechanisms underlying the state-dependent in-service times seen in Figure 1d. Some of the mechanisms are suggested by prior literature and one is original to this study.

5. Slowdown: Queues within the Queue

To identify mechanisms that impact the speed of ED service, we shift from a patient point of view to a server point of view. This change in point of view is necessary because serving a patient in the ED is not a single task but rather a collection of many tasks performed by many resources, each of which can be thought of as an individual server with its own queue of tasks. We focus first on Slowdown, or mechanisms that increase a patient's in-service time. Prior literature has shown evidence of a link between service times and server busyness in healthcare settings. Kuntz et al. (2014) shows evidence of hospital inpatient length of stay increasing with moderate increases in system load, and Armony et al. (2013) shows ED service times increasing with ED census. Similarly, Berry Jaeker and Tucker (2013) reports evidence of inpatient length of stay being a function of workload both

in a given hospital unit and in related units that share resources. None of these papers attempt to identify the mechanisms causing the relationship between load and length of stay, and they do not make use of resource level data as we do.

Two mechanisms that have been shown to lead to Slowdown of individual servers are fatigue and multitasking. For example, several studies in medical and ergonomics journals have shown that fatigue leads to diminished productivity (e.g., Setyawati 1995, Caldwell 2001). Similarly, Kc and Terwiesch (2009) finds that fatigue caused by extended periods of high workload leads to decreased productivity in both hospital transportation and cardiac ICU care. Kc (2013) examines the effect of ED physician multitasking on service time and finds that moderate levels of multitasking improves productivity, but eventually increased multitasking leads to decreasing throughput of patients. Kc (2013) posits that coordination waits (queuing of tasks) and cognitive switching costs are responsible for the throughput decay.[5]

In the analytical queuing literature, multitasking is equivalent to queues with shared processors (e.g., Yamazaki and Sakasegawa 1987, Aksin and Harker 2001). Shared processor models assume that a server processes multiple customers simultaneously and splits its processing capacity across all items in service. This leads to service times increasing as the number of customers in service increases. For example, Aksin and Harker (2001) models a multi-server call center with multiple customer classes and a single shared information management system that slows down as it performs more simultaneous operations. The key finding is that the system throughput decay caused by processor sharing is a function of both the offered load on the system and the proportion of a customer's service that requires use of the shared resource. This is relevant for our ED setting since many resources in the ED are shared resources (e.g., nurses, doctors, equipment) and EDs regularly operate under high offered loads.

When considering human servers, such as care providers in the ED, the term "multitasking" is a misnomer. These workers are rarely serving multiple customers simultaneously; the nurse is only at one bedside at a time. Rather, while care providers may have responsibility for several patients simultaneously, they perform tasks for patients individually and sequentially (e.g. draw blood from Patient X, then give medicine to Patient Y). In this sense, the human server is better conceptualized as a "queue within the queue." For example, a nurse can be viewed as a very flexible server that can perform varied tasks such as taking vital signs, delivery medication, or drawing a specimen for a lab test. These tasks arrive in the individual nurse's virtual work queue as doctors write orders or as patients' needs arise. This arrival rate is presumably higher when the nurse is responsible for

[5] Another possible cause of Slowdown is interruptions. Chisholm et al. (2000) documents that interruptions are a regular occurrence for ED care providers, and Dobson et al. (2013) provide analytical results showing that interruptions can lead to delays in ED patient care. Our data, does not allow us to observe interrruptions in the ED.

more patients (higher multitasking, in the language of Kc (2013)). By Little's Law, the higher the arrival rate of tasks, the longer it takes the average task to flow through the nurse's work queue, all else equal. This increased task flow time leads to Slowdown, or longer in-service times for the patients that are waiting on the individuals tasks.

5.1. Model Formulation

We are interested in determining if higher workload leads to longer completion times for activities performed by individual servers. We focus here on nursing tasks, but similar analysis also applies to other resources such as physicians and x-ray machines. Our data allows us to isolate two nursing specific tasks: medication delivery and lab specimen collection. As mentioned is Section 2, medication and lab orders are initiated when a physician enters the order in the electronic medical record. The order becomes part of a nurse's virtual or mental queue of tasks and eventually is carried out and marked as complete in the electronic medical record. The duration between the order creation and the order completion is what we take as the order completion duration and includes both the time the order spends in the nurse's virtual queue and the activity time spent carrying out the order.

Ideally we would analyze the workload and order completion duration for each nurse individually, however the data does not contain sufficient information to do so. Rather, we pool the data across the ED for each hour of the sample and relate the mean order completion duration in hour h to the workload of hour h. Since we are again dealing with durations, we use an accelerated-failure-time model similar to Equation 1.

$$ln(MeanDuration_h) = \alpha + \beta_1 MedOrders_h + \beta_2 LabOrders_h + \beta_3 \overline{BedCensus_h} + \mathbf{Y_h}\boldsymbol{\delta} + \mathbf{Z}_h\boldsymbol{\phi} + \epsilon_h \tag{2}$$

The index h denotes an hour in the sample. We estimate Equation 2 for two definitions of $MeanDuration$:

1. Mean medication order completion time in hour h (in minutes)
2. Mean lab sample collection completion time in hour h (in minutes)

The variables $MedOrders_h$ and $LabOrders_h$ are the number of medication and lab orders written in hour h respectively. $\overline{BedCensus_h}$ is the mean number of ED beds in use in hour h. Positive coefficients on these variables indicates order completion duration increasing with workload. Because order completion duration is likely impacted not only by contemporaneous workload, but also by work from prior periods, we include lags of the dependent variable ($\mathbf{Y_h}$). A Durbin alternative test for serial correlation (Durbin 1970) shows that including five lags of the dependent variable is sufficient to remove autocorrelation of the error terms (thus the model is an AR(5) model). $\boldsymbol{Z}$ is a vector of time-related control variables including year, month, hour of day, weekend indicator,

and the interaction of hour of day and weekend. Because the dependent variables are estimated means (the mean order completion duration of all orders written in hour h), we use weighted least squares to estimate the model with the weights set equal to the number of orders ordered in hour h (Wooldridge 2009). We assume ϵ_h to be normally distributed.

5.2. Results

The estimation results of the order completion duration models are shown in Table 4. Because the

Table 4 Order completion duration of nurse-specific tasks

	(1) Medication Delivery Time	(2) Lab Collection Time
Medication Orders	0.029***	0.002
	(0.002)	(0.002)
Lab Orders	0.006***	0.005***
	(0.001)	(0.000)
ED Bed Census	0.008***	0.007***
	(0.001)	(0.001)
N	17,579	20,436

Robust standard errors in parentheses
5 lags of the dependent variable included
Controls not shown: year, month, weekend × hour
* $p<0.10$, ** $p<0.05$, *** $p<0.01$

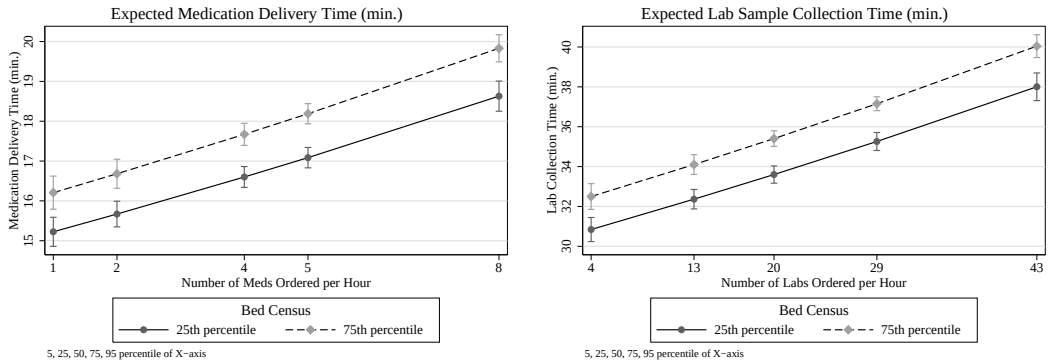
dependent variable is log-transformed, the coefficient estimates can be interpreted as percentage changes of the dependent variable. For example, Model 1 shows that an additional medication order is associated with a 2.9% increase in mean medication delivery time in the present time period. For medication delivery, all three workload measures are positive and significant leading to increased order completion duration. For lab collection, the number of new medication orders does not have a significant impact. This could be caused by nurses prioritizing lab orders over delivering medications.

Figure 2 displays plots of expected durations predicted by the models. Both subplots show order completion duration increasing approximately linearly with the number of orders written, suggesting that basic queuing effects are driving this increase. The plots also show that increased bed census, an indirect measure of nurse workload, also leads to increased order completion time.

We perform similar analyses on other resource-specific tasks (e.g. time between a patient being placed in a bed and first seeing a physician, radiology scan completion time), and find similar results. This server-level analysis shows that tasks performed by individual servers take longer to be completed when the workload is high due to queuing delays at the level of the server. These delays create the queue within the queue effect and act to increase the in-service time of the patient.

Figure 2 Order completion duration as a function of workload

(a) Medication Delivery Duration (b) Lab Collection Duration



Note: Error bars indicate 95% confidence interval of the mean.

6. Speedup: Rushing, Cutting, & Shifting

We now examine mechanisms that could cause Speedup, that is a reduction of patient in-service durations. The subset of queuing theory focused on optimal control of queues provides theoretical motivation for Speedup behavior. Dynamic control queues dynamically adjust to system state parameters such as the queue length. Going back to Crabill (1972), several papers have explored optimal control policies that minimize average cost per unit time by adjusting the service time, and have proven under increasingly weaker assumptions the existence of an optimal service time policy that is monotone decreasing in queue length (e.g., Stidham and Weber 1989, George and Harrison 2001). The intuition behind such a policy is based on the assumptions that the system waiting cost per unit time increases with queue length and that there is a cost to decreased service time, either in terms of labor, effort, or reduced quality. Thus, as the queue length grows, the waiting costs eventually outweigh the cost of faster service and the optimal response is to speed up the service time. We examine three such Speedup mechanisms: rushing, task reduction, and early task initiation.

6.1. Rushing: Nurse-Patient Interaction Times

Perhaps the simplest form of service time reduction is *rushing*. That is, a server simply working faster as the workload increases. Schultz et al. (1998) finds this sort of acceleration behavior in a lab experiment, and Kc and Terwiesch (2009) is the first paper to show this behavior in the field. It finds that hospital transporters work faster when the workload is high. Similarly, Tan and Netessine (2012) and Staats and Gino (2012) find evidence of rushing under load with restaurant waiters and loan application processors, respectively.

The results in Section 5.2 show that as workload increases it takes nurses longer to deliver medication and collect lab samples. However, this effect is consistent with a Little's Law effect rather

than a change in actual working pace of the nurse. Stated differently, the completion duration increases because the task spends more time waiting in the nurse's virtual queue, not because the nurse necessarily spends any more time performing the task. The data do not allow us to observe actual time spent delivering a medication or collecting a lab sample. Thus we need different data to separate queue time from service time and shed light on actual time spent serving a patient.

While it is the physician who decides what should be done for a patient, it is largely the responsibility of nurses to deliver service at the bedside. These bedside service tasks may be direct, such as taking vitals signs, administering medications, or talking with the patient about her condition and treatment, or indirect, such as assisting a physician performing a procedure. The nurse has limited discretion in what bedside service must be delivered (e.g. vital signs must be taken periodically, physician orders must be completed, etc.), but the nurse does have discretion over the speed at which tasks are accomplished. Therefore, the amount of time a nurse spends in the patient room is a measure of task speed. It is possible that nurses rush and spend less time in direct patient care when the ED is busy.

6.1.1. Model Formulation As described in Section 3, data from the nurse communication and tracking system allow us to observe the number of nurse-patient interactions, the duration of each interaction, and an identifier of the nurse involved in the interaction. We are interested in examining how nurse-patient interactions change as a function of the system busyness. We estimate models of the form

$$ln(y_i) = \alpha + \beta_1 Census_i + \beta_2 Census_i^2 + \mathbf{W}_i\boldsymbol{\theta} + \mathbf{Z}_i\boldsymbol{\phi} + \epsilon_i \tag{3}$$

where the index i denotes a patient visit to the ED. We estimate Equation 3 for three dependent variables (y_i):

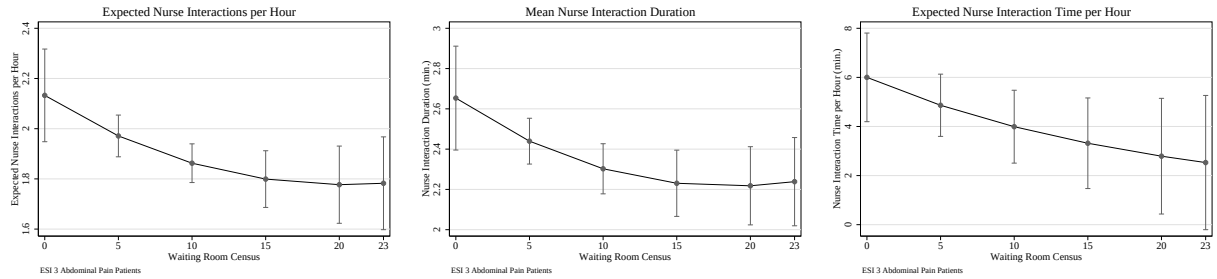
1. Number of nurse-patient interactions per patient hour
2. Duration of individual nurse-patient interactions (in minutes)
3. Total nurse-patient interaction duration per patient hour (in minutes)

$\mathbf{W}_i$ is a vector of patient-visit specific covariates including age, gender, triage level, physician, and the number of physician-ordered diagnostics. We include the number of physician-ordered diagnostics to control for potential changes in the number of tasks required for each patient (Section 6.2 explores this topic in detail). $\boldsymbol{Z}_i$ is a vector of time-related control variables including month, a weekend indicator, and hour of day. We assume ϵ_i is normally distributed.

We use the waiting room census at the time the In-Service Phase begins as the *Census* independent variable. We focus on the waiting room census rather than the total or bed census for two

Figure 3 Duration elements as a function of census

(a) Interaction per Patient Hour (b) Duration of a Single Interaction (c) Nurse Interaction Time per Hour



Note: Error bars indicate 95% confidence interval of the mean.

reasons. First, there is much more variation in the waiting room census, with a coefficient of variation of 0.76, as compared to 0.26 and 0.19 for total census and ED bed census, respectively. Second, the waiting room census potentially provides a more interesting insight since it has no impact on the nurse's immediate workload. Nurses can easily observe the waiting room census on electronic dashboards around the ED. We include both the linear and quadratic forms of the census variable to allow for a more flexible functional form.

Table 5 Nurse-patient interaction models

	(1)	(2)	(3)
	Interactions per Hour	Duration per Interaction	Interaction Duration per Hour
Wait Census	-0.018***	-0.019**	-0.043*
	(0.007)	(0.009)	(0.009)
Wait Census^2	0.000**	0.001*	0.000
	(0.000)	(0.000)	(0.001)
Marg. Effect			
Wait Census	-0.021**	-0.027**	-0.200***
N	1,781	1,781	1,781

Robust standard errors in parentheses
Marginal effect estimated at median census: 8 patients
Controls not shown: age, gender, triage level, physician, # of orders, month, weekend, hour
* $p<0.10$, ** $p<0.05$, *** $p<0.01$

6.1.2. Results The results from estimating Equation 3 are shown in Table 5 and plots of predicted values are shown in Figure 3. All three models show marginal effects that indicate a negative response to waiting room census. As the waiting room census increases, the number of interactions per patient hour decreases and the mean duration of those interactions also decreases leading to a decrease in the total nurse-patient interaction time per patient hour. Figure 3c shows that for ESI 3 patients, nurse-patient interaction time drops by approximately 58% (from 6.0 minutes

to 2.5 minutes) as the waiting room census ranges from the 5th percentile (0 people) to the 95th percentile (23 people).

This result is particularly interesting since the waiting room census has no direct impact on nurse workload. Nurses are assigned a predetermined block of rooms and once those rooms are full the workload does not change with changes in the waiting room census. It appears that the nurses instinctively react to the waiting room census by rushing in an attempt to speed up the patient service. However, as evidenced by the findings in Section 5.2, this Speedup effort is not sufficient to outweigh the Slowdown effect caused by queuing delays, at least with regard to medication delivery and lab specimen collection. Further, the positive coefficient on the quadratic term of census in Model 3 and the shape of the curve in Figure 3c show that the marginal amount of nurse rushing diminishes (in magnitude) as the waiting room census increases. Thus, nurse rushing alone is not sufficient to cause the decrease in in-service times seen in Figure 1d.

6.2. Task Reduction & Early Task Initiation

Papers by Hopp et al. (2007) and by Alizamir et al. (2013) build on the optimal queue control literature mentioned above and suggest another Speedup mechanism; *task reduction*. Hopp et al. (2007) describes a service system with discretionary length service tasks that are concave-increasing in value with time. A holding cost is incurred per unit time for each customer in the system. This leads to an optimal service policy that sets a service cutoff time for every value of queue length. This policy is monotone decreasing in queue length. Similarly, Alizamir et al. (2013) models a diagnostic service as a stochastic sequence of diagnostic tests. Each test informs the server's probability estimation of the customer's type. This specification can lead to an optimal policy that sets a maximum number of tests for each queue length. This maximum is decreasing in queue length. The common element of these papers is that it is a change in the service content, not the service rate, which leads to a change in the service time per customer. Oliva and Sterman (2001), Kc and Terwiesch (2009), and Chan et al. (2011) are all suggestive of this sort of task reduction based Speedup which Oliva and Sterman (2001) colloquially refers to as "cutting corners." (We prefer the term "task reduction" because "cutting corners" implies that a reduction in tasks is undesirable. However, as described in Debo et al. (2008), it may be that unnecessary tests are performed when the ED is lightly loaded and that testing is reduced to its clinically appropriate level when the ED is busy.)

In the Hopp et al. (2007) model, the variable under the server's control is service time itself. In the ED, the physician is the key decision maker and can indirectly control service time by varying the quantity of diagnostic tests ordered. Diagnostic tests (labs and scans) generally increase service time since they require time to perform and process. Further, diagnostic tests draw on various

resources (nurses, lab technicians, transporters, radiology technicians, radiologists, etc.) which are all susceptible to the queue within the queue Slowdown effects examined in Section 5. Therefore, a physician that is trying to speed up the ED might choose to order fewer diagnostic tests. Further, the triage nurse also has the option of ordering diagnostic tests and might also exhibit task reduction behavior.

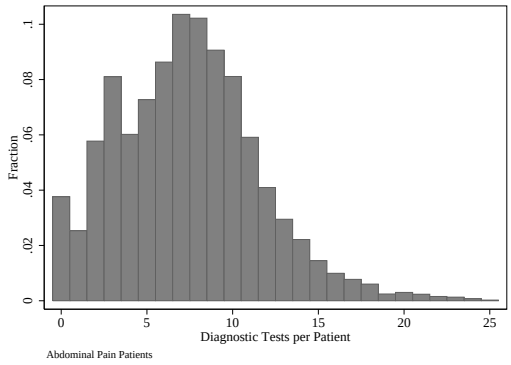
While task reduction is a Speedup mechanism that can be implemented by a single server, we propose the mechanism of *early task initiation* as a Speedup mechanism that may exist between servers. The ED process is equivalent to a tandem queue with triage being the first stage and service in the treatment room (the In-Service Phase) being the second stage. The Waiting Room Phase is the buffer between these two stages. If the triage nurse is able to anticipate what tests will be ordered by the physician, she can order these tests at triage and they will be processed (or at least started) while the patient is waiting in the waiting room. Then when the patient enters the In-Service Phase and is seen by the physician the tests will already be under way or may even be ready for review, thus reducing in-service time. The benefit of early task initiation grows when the waiting room census is high and waiting times are long since this allows even long diagnostic tests to be performed while the patient is waiting, shifting even more time out of the In-Service Phase.

However, a downside of early task initiation is that the nurse may be "placing bets," in that the nurse may not be certain what tests the doctor will want and may order unneeded tests. This could be due to the nurse having less training and skill than the doctor, or due to the limited information available from a triage examination, or simply due to the difficulty of one person trying to guess another's actions. Over-testing is undesirable because it increases monetary costs, medical risk for the patient (if the test is risky), and workload on the diagnostic resources. Since the risk of overtesting does not change with the census level while the benefit increases with census, it is likely that early task initiation increases with census.

6.2.1. Model Formulation To examine these mechanisms, we are interested in estimating how the count of diagnostic orders per patient changes in relation to the census. Since the dependent variable is discrete and small we need to use a count-type model rather than a continuous-variable regression model. Further, as seen in Figure 4, the excess of zero counts suggests the need for a zero-inflated model. We use a zero-inflated negative binomial (ZINB) model for all of these studies (Hilbe 2011). The ZINB model combines a binary logit process with probability density $f_1(\cdot)$ and a negative binomial count process with probability density $f_2(\cdot)$ to create the combined density

$$f(y|\mathbf{x}) = \begin{cases} f_1(1|\mathbf{x}_1) + \{1 - f_1(1|\mathbf{x}_1)\} f_2(0|\mathbf{x}_2) & \text{if } y = 0 \\ \{1 - f_1(1|\mathbf{x}_1)\} f_2(y|\mathbf{x}_2) & \text{if } y \geq 1 \end{cases} \tag{4}$$

Figure 4 Distribution of diagnostic tests ordered for abdominal pain patients



Note that this formulation is somewhat counterintuitive (albeit standard practice) in that a "success" of the binary process corresponds to $y = 0$, whereas a "failure" corresponds to y being determined by a negative binomial count process. This model has the conditional mean

$$E[y|\mathbf{x}] = \frac{1}{1+\exp(\mathbf{x_1}\boldsymbol{\eta_1})} \times \exp(\mathbf{x_2}\boldsymbol{\eta_2}) \tag{5}$$

We estimate four versions of Equation 5:

1. $y =$ Count of diagnostics ordered by physician; not controlling for orders placed at triage
2. $y =$ Count of diagnostics ordered by physician; controlling for orders placed at triage
3. $y =$ Count of diagnostics ordered at triage
4. $y =$ Count of total diagnostics ordered (triage nurse + physician)

In Equations 4 and 5 the vectors $\mathbf{x_1}$ and $\mathbf{x_2}$ represent the covariates to be included in the regression model and the vectors $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are the parameters to be estimated. The covariate vectors $\mathbf{x_1}$ and $\mathbf{x_2}$ need not be the same, but for our purposes they are the same unless noted otherwise on the result table. The parameter vectors $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are estimated jointly by maximum likelihood using the log-likelihood function shown in the Appendix A. For $\boldsymbol{\eta}_1$, a positive coefficient indicates a decrease in the expectation of the dependent variable with an increase in the given independent variable, while the opposite is true for $\boldsymbol{\eta}_2$.

For models 1, 3, and 4 we formulate the linear predictors $\mathbf{x}_{i,1}\boldsymbol{\eta}_1$ and $\mathbf{x}_{i,2}\boldsymbol{\eta}_2$ as follows:

$$\mathbf{x}_{,j}\boldsymbol{\eta}_j = \alpha_j + \eta_{j,1}Census + \eta_{j,2}Census^2 + \mathbf{W}_j\boldsymbol{\theta}_j + \mathbf{Z}_j\boldsymbol{\phi}_j \text{ for } j = 1,2 \tag{6}$$

$\mathbf{W}_j$ is a vector of patient-visit specific covariates such as age, gender, race, triage level, and physician. $\mathbf{Z}_j$ is a vector of time related control variables such as year, month, shift, and a weekend indicator variable.[6] For model 2, Equation 6 is modified to also include the number of diagnostics ordered at triage.

[6] The shift variable indicates the three main physician work shifts: 7:00am-3:00pm, 3:00pm-11:00pm, and 11:00pm-7:00am. We use this shift indicator rather than an hour of day indicator because it captures much of the time of day effect with only two dummy variables rather than twenty three.

Similar to Equation 3, we use the waiting room census as the census measure. The waiting room census is directly observed by triage nurses because the triage rooms are adjacent to the waiting room. The waiting room census is readily viewed by physicians on electronic dashboards around the ED. Anecdotal evidence and research team observation suggests that ED nurses and physicians focus on the waiting room census as a key indicator of the busyness of the ED. For models 1, 2 & 4, the census is measured at the time the In-Service Phase begins. For Model 3, census is measured at the time of arrival to the ED. We include both the linear and quadratic forms of the census variable to allow for a more flexible functional form.

Table 6 Impact of waiting room census on the number of diagnostic tests ordered

	(1) Doctor Orders	(2) Doctor Orders	(3) Triage Orders	(4) Triage & Doctor Orders
Inflate ($\mathbf{x_1}\boldsymbol{\eta_1}$)				
Wait Census	0.018	-0.002	-0.234***	-0.047 *
	(0.021)	(0.021)	(0.013)	(0.026)
Wait Census^2	0.000	0.000	0.006***	0.001
	(0.001)	(0.001)	(0.000)	(0.001)
Triage Orders		0.228***		
		(0.028)		
Neg.Bin. ($\mathbf{x_2}\boldsymbol{\eta_2}$)				
Wait Census	-0.009***	-0.002	0.023***	0.000
	(0.002)	(0.002)	(0.007)	(0.002)
Wait Census^2	0.000**	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Triage Orders		-0.140***		
		(0.006)		
Marg. Effect				
Wait Census	-0.044***	0.000	0.040***	0.013*
Triage Orders		-1.01***		
N	12,288	12,288	13,691	12,288

Robust standard errors in parentheses
Controls not shown: age, gender, race, triage level, physician, year, month, weekend × shift
* $p<0.10$, ** $p<0.05$, *** $p<0.01$

Note: Results are from zero-inflated negative binomial models. The *Inflate* coefficients are for the zero-inflation part of the model and the *Neg.Bin.* coefficients are for the negative binomial part of the model. The marginal effects shown are the mean marginal effect over all observations.

6.2.2. Results Table 6 shows the results of estimating the four models described above. The top panel shows the coefficients estimated for the zero-inflation part of the model ($\boldsymbol{x_1}\boldsymbol{\eta_1}$). Positive coefficients indicate an increase in the probability of a zero outcome, which reduces the mean outcome. The middle panel shows the coefficients estimated for the negative binomial part of the model ($\boldsymbol{x_2}\boldsymbol{\eta_2}$). Positive coefficients indicate an increase in the mean outcome. Due to the non-linear,

and two-part nature of the ZINB model, direct interpretation of the coefficients is difficult. The bottom panel reports the mean marginal effect of the variables of interest.

Model 1 shows that doctors order fewer diagnostics as census increases with a mean decrease of 0.044 diagnostics per one person increase in waiting room census. Figure 5a shows mean doctor orders dropping from 7.1 orders per patient to about 6.2 orders per patient as the waiting room census ranges from the 5th percentile (0 people) to the 95th percentile (23 people). This result would seem to show that doctors are engaging in task reduction when the ED becomes crowded. However, once we control for the number of tests ordered at triage (Model 2), we see that the waiting room census no longer has significant coefficients or marginal effect. In other words, there is no evidence of crowd-induced task reduction or "corner cutting" by the doctors. Instead, the number of triage orders becomes the significant predictor. Interestingly, the mean marginal effect is approximately negative one suggesting that early task initiation is happening. The mean number of doctor-ordered tests drop by about one for each additional triage test ordered. Stated differently, this result indicates a fairly efficient shifting of diagnostic ordering from doctor to triage nurse. However, the 95% confidence interval on the marginal effect ranges from -1.7 to -0.4, so the work shifting might not be perfectly one-for-one.

Model 3 provides further evidence of early task initiation rather task reduction with triage nurses increasing diagnostic test orders by 0.040 tests per patient on average for each additional person in the waiting room. Figure 5b shows mean triage diagnostic orders increasing from 0.2 to almost 1.2 orders per patient as the waiting room census varies over its typical range.

Model 4 illustrates the drawback of early task initiation. Model 4 shows that total diagnostic testing (triage orders plus doctor orders) increases slightly as waiting room census increases (Figure 5c). However, this result is only marginally significant (p-value of 0.071). This is evidence that triage nurses do not perfectly guess which tests doctors will want and they sometimes order unnecessary tests leading to an increase in total testing. Since overtesting is costly, early task initiation should only be used when the in-service time savings warrant its use.

To get a sense of the impact on in-service time of these changes in testing, we re-estimate Model 4 of Table 3 (Section 4.2) and add in variables for the number of diagnostics ordered at triage and by the doctor. Both variables are significant predictors of in-service time with mean marginal effects of 0.21 and 0.38 hours respectively. Thus, shifting one diagnostic test from being doctor ordered to being triage ordered has a net effect of reducing mean in-service time by 0.17 hours (~10 minutes). This benefit may be eroded, however, by a slight rise in total testing.

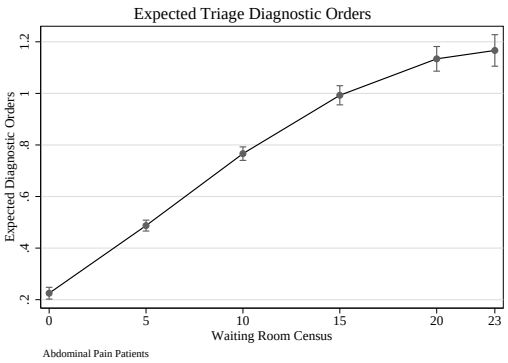
7. Robustness to Endogenous Treatment and Selection

As with all empirical studies, we must give thought to potential endogeneity issues. There are two potential sources of endogeneity bias in our study of doctor-ordered tests: triage-ordered testing

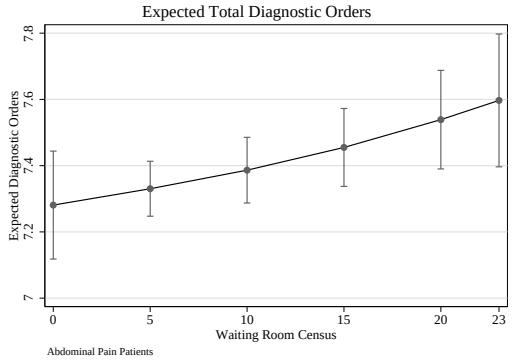
Figure 5 Expected number of diagnostic tests ordered at triage and by the physician

(a) Model 1: Doctor Ordered Diagnostics

(b) Model 3: Triage Diagnostic Orders



(c) Model 4: Total Diagnostic Orders



Note: Error bars indicate 95% confidence interval of the mean.

and patient abandonment. Triage-ordered diagnostic testing is not randomly assigned, but rather is a decision made by a triage nurse based on the characteristics of the patient, some of which are observed (e.g., age, gender, race) and some of which are unobserved by the researcher (e.g., countenance, sweating, pallor). Similarly, the doctor's decision to order diagnostics is likely driven by many of the same observed and unobserved patient characteristics. A shared unobserved variable could induce correlation in the triage testing and doctor testing models leading to biased estimates of the coefficients.

The issue of patient abandonment, also known as Left Without Being Seen (LWBS), further complicates the issue. Patients sometimes abandon the queue after being triaged but before being seen by a doctor. This abandonment filters the population that the doctor sees. If this filtering changes with crowding, then the doctor is seeing a different patient mix during times of high and low crowding. Further, this filtering is a potential problem because the abandonment rate is affected by whether diagnostics are ordered at triage and is possibly driven by the same unobservable covariates affecting triage-ordered testing and doctor-ordered testing. Thus, there is the potential for a

three-way interaction between triage-ordered testing, abandonment, and doctor-ordered testing. For example, a patient with abdominal pain who is pale and sweaty may have an increased probability of receiving diagnostic tests ordered both in triage and by the doctor, and might be highly likely to wait to be served since the patient feels quite sick. This would lead to positive correlations among the three equations. If there is a bias, it is likely that the bias is toward sicker patients remaining and receiving more tests during high crowds. Note, however, that all these potential issues only become problematic if the observed covariates are not rich enough to capture the differences between patients.

The "ideal" test for endogeneity would be a three-equation model that simultaneously estimates the endogenous treatment (triage-ordered testing), the self-selection (abandonment), the resulting zero-inflated count outcome (doctor-ordered testing) and the three pairwise correlations. Unfortunately, to the best of our knowledge, no such model exists. The closest model we are aware of is the sample-selection-endogenous-treatment model from Bratti and Miranda (2011). However, this model uses a Poisson model for the final outcome and generally fails to converge with our overdispersed and zero-inflated data. In lieu of an ideal test, we present several pieces of supporting information that point to the conclusion that our results are robust to the potential endogeneity problems.

We begin with the patient abandonment issue and the sample selection problem it potentially creates. Overall, 9.5% of observed abdominal pain patients abandon the queue. However, the rate ranges from less than 1% under low waiting room census to 20% under high waiting room census. We use a Heckman-style bivariate probit selection correction model to test for unobserved correlation between patient abandonment and doctor testing (de Ven and Praag 1981, Greene 2012). We treat both the abandonment decision and doctor testing as binary outcomes and formulate the model as follows:

$$\begin{gathered} S^* = \alpha_1 + \beta_{1,1} WaitCensus_{arrival} + \beta_{1,2} WaitCensus^2_{arrival} + \beta_{1,3} \mathbf{1}(TriageTests > 0) + \mathbf{W}\boldsymbol{\theta_1} + \mathbf{Z}\boldsymbol{\phi_1} + \varepsilon_1 \\ STAY = 1 \text{ if } S^* > 0,\ 0 \text{ otherwise} \end{gathered} \tag{7}$$

$$\begin{gathered} D^* = \alpha_2 + \beta_{2,1} WaitCensus_{service} + \beta_{2,2} WaitCensus^2_{service} + \beta_{2,3} TriageTests + \mathbf{W}\boldsymbol{\theta_2} + \mathbf{Z}\boldsymbol{\phi_2} + \varepsilon_2 \\ \mathbf{1}(DoctorTests > 0) = 1 \text{ if } D^* > 0,\ 0 \text{ otherwise} \end{gathered} \tag{8}$$

The variable $WaitCensus_{arrival}$ is the waiting room census at the time of a patient's arrival and the variable $WaitCensus_{service}$ is the waiting room census at the time a patient starts the in-service phase of the visit (consistent with the census variables used in Section 6.2.2). We note that the absence of $WaitCensus_{arrival}$ in the second equation provides an exclusion restriction

that helps with model identification even though the model is technically identified by the non-linearity of the probit equations. The variable $TriageTests$ is the number of diagnostic tests ordered at triage and $DoctorTests$ is the number of tests order by the doctor. The vector $\mathbf{W}$ contains the patient covariates age, gender, race, and triage level. The vector $\mathbf{Z}$ contains controls for year, month, weekend, and shift. ε_1 and ε_2 are assumed to be standard bivariate normally distributed with correlation coefficient ρ, and Equation 8 is only observed if $STAY = 1$ (the patient does not abandon). If $\rho = 0$, this indicates that the control variables are adequately controlling for the selected sample and the models can be estimated separately without significant bias.

Table 7 Heckman and bivariate probit models to test for endogeneity

	(1)	(2)	(3)
First Stage	Stay (Y/N)	Triage Test (Y/N)	Triage Test (Y/N)
Wait Census	-0.196***	0.120***	0.134***
	(0.010)	(0.005)	(0.006)
Wait Census^2	-0.004***	-0.003***	-0.003***
	(0.000)	(0.000)	(0.000)
Second Stage	Doctor Test (Y/N)	Stay (Y/N)	Doctor Test (Y/N)
Wait Census	0.004	-0.184 ***	0.001
	(0.009)	(0.015)	(0.008)
Wait Census^2	0.000	0.004 ***	0.000
	(0.000)	(0.000)	(0.000)
ρ	-0.128	0.188	-0.015
	(0.159)	(0.132)	(0.032)
Age, Race, Gender, Triage	Yes	Yes	Yes
Year, Month, Weekend, Shift	Yes	Yes†	Yes
Triage Orders	Yes		Yes††
N	13,571	13,571	12,288

Robust standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

†Shift variable excluded from second stage equation

††Variable included in Doctor Test portion of model only

The estimation results for Equations 7 and 8 are shown in Model 1 of Table 7. The coefficients in the upper panel show that the probability of staying (not abandoning) decreases with load, as one would expect. The coefficients in the lower panel indicate that doctor-ordered testing is not affected by census level when we control for the number of triage-ordered tests. This is consistent with the results from Model 2 of Table 6. Most importantly, we see that the estimated correlation between the two equations is not significantly different from zero and therefore we need not be overly concerned about patient abandonment creating a selection bias.

We next check for unobserved correlation between triage-ordered testing and patient abandonment. We use a bivariate probit model which is conceptually similar to the selection model above,

but without needing to adjust for the selected sample of the second equation. The equations are as follows:

$$\begin{gathered} T^* = \alpha_1 + \beta_{1,1} WaitCensus_{arrival} + \beta_{1,2} WaitCensus^2_{arrival} + + \mathbf{W_1}\boldsymbol{\theta_1} + \mathbf{Z}\boldsymbol{\phi_1} + \varepsilon_1 \\ \mathbf{1}(TriageTests > 0) = 1 \text{ if } T^* > 0,\ 0 \text{ otherwise} \end{gathered} \tag{9}$$

$$\begin{gathered} S^* = \alpha_2 + \beta_{2,1} WaitCensus_{arrival} + \beta_{2,2} WaitCensus^2_{arrival} + \beta_{2,3}\mathbf{1}(TriageTests > 0) + \mathbf{W_2}\boldsymbol{\theta_2} + \mathbf{Z}\boldsymbol{\phi_2} + \varepsilon_2 \\ STAY = 1 \text{ if } S^* > 0,\ 0 \text{ otherwise} \end{gathered} \tag{10}$$

The variables are all defined as in Equations 7 and 8 with the exception that the shift indicator variable is drop from $\mathbf{W}_2$ to provide an exclusion restriction.

The estimation results for Equations 9 and 10 are shown in Model 2 of Table 7. The results show that the census coefficients are all significant and in the direction we expect; triage testing and abandonment (not staying) both increase with crowding. As with Model 1, the between-equation correlation is insignificant indicating that a triage-ordered testing treatment bias is not an issue.

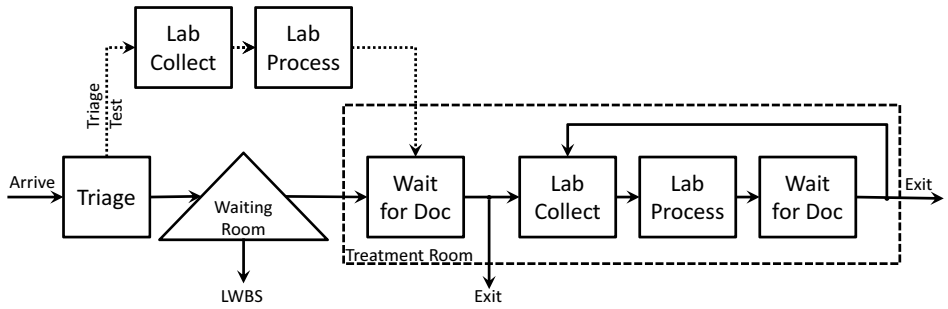
We lastly examine the potential endogeneity between triage-ordered testing and doctor-ordered testing and we again use a bivariate probit model similar to Equations 9 and 10. We ignore the middle step of abandonment based on the above results showing that abandonment is not creating a significant bias. The results of this analysis are shown in Model 3 of Table 7. The coefficients in the upper panel are as expected indicating increased triage-ordered testing with increased crowding, and the point estimates are similar to those of Model 2. The coefficients in the lower panel are also as expected, showing no change in doctor-ordered testing with census, controlling for triage testing. The correlation coefficient is insignificant indicating the biprobit model is not necessary.

To check the robustness of our findings regarding the presence of task reduction (or the lack thereof), we re-estimate Model 1 from Table 6 with a special subset of the data. We test for changes in the number of doctor-ordered tests as a function of load for patients that receive no triage-ordered tests. Clearly, this is a non-random sample, but it is free of any convoluting effects of doctors responding to triage-ordered testing. We find that there is no evidence of the number of tests ordered by doctors changing with load, just as we concluded in Section 6.2.2.

8. Simulation

Given our findings of several mechanisms that generate state-dependent service times in the ED, we are interested in determining what impact these have on system performance models. To estimate the impact of the state-dependencies, we build a discrete-event-simulation model of the ED. Figure 6 diagrams the patient flow in the model. While the model is abstracted from reality, we maintain the essential elements that generate Slowdown and Speedup. Lab Collect and Wait for Doc durations increase with waiting room census leading to Slowdown. Speedup is generated through early

Figure 6 Patient Flow in Simulation Model



task initiation whereby triage-ordered testing increases and doctor-ordered testing decreases with load.[7] One additional state-dependency included in the model is the Left Without Being Seen or abandonment rate. While we do not focus on this phenomenon in this paper, the results of Models 1 and 2 in Table 7 clearly show a strong positive correlation between the waiting room census and patients abandoning the queue (we refer the reader to Batt and Terwiesch (2014) for complete study of patient abandonment behavior in the ED).

We test three configurations of the model (Table 8). In the first configuration (Model 1), all state-dependencies are active and the model is tuned to match the average performance of our study ED. In the second configuration (Model 2), only the Speedup and Slowdown state-dependencies are deactivated by fixing the variables at their mean values. In the third configuration (Model 3), all state-dependencies, including LWBS, are deactivated. The simulation is run for 50,000 simulated hours and standard errors are calculated using the batch-means process with batches of length 200 hours (Law 2007).

Table 8 Simulation Results

Outcome (mean)	(1) State-Dependent	(2) State-Independent (except LWBS)	(3) State-Independent (incl. LWBS)
Queue Length	8.3 (0.21)	8.8 (0.17)	9.9 (0.64)
Wait Time (hr.)	1.6 (0.04)	1.7 (0.03)	2.0(0.1)
Length of Stay (hr.)	5.6 (0.05)	5.8 (0.03)	6.1 (0.10)
LWBS %	5.8% (0.002)	6.2% (0.001)	8.6% (0.001)

Standard error in parentheses

Comparing Model 2 to Model 1 we see that ignoring the Speedup and Slowdown mechanisms leads to a modest overestimation of all of the system performance measures. Comparing Model

[7] We leave the lab processing time distribution stationary because the lab serves the entire hospital and the ED demand has little impact on lab times.

3 to Model 1 we see that ignoring all state-dependencies leads to a larger overestimation of all performance measures. This potential overestimation is managerially relevant since similar models are commonly used for hospital staffing and planning purposes. These planning models are becoming increasingly important as the Centers of Medicare and Medicaid Services (CMS) begin to phase in new ED reporting guidelines and performance targets. Hospitals are beginning to be required to report ED performance measures such as median wait time, median length of stay, and Left Without Being Seen percentage (Centers for Medicare & Medicaid Services 2012). Eventually, target values will be established and hospitals will be reimbursed based on their performance relative to the targets. Thus, a hospital that is making planning decisions based on a model which does not include the identified state-dependencies is likely to overinvest in resources and staffing to meet the CMS targets.

9. Discussion & Conclusion

Prior research has shown that many service systems exhibit state-dependent behavior with service times that change in relation to the busyness of the system. In this paper we explore several mechanisms that lead to state-dependent service times. We find evidence of both Speedup and Slowdown mechanisms.

We find that the time a patient spends in the In-Service Phase of an emergency department visit exhibits an inverted-U relationship with the level of crowding in the waiting room. We show that the slowdown is at least partly due to "queue within the queue" effects. This occurs when tasks (e.g. medication orders) must wait in a virtual work queue of individual servers or resources. As the workload increases, the total time for a task or order to work its way through the queue and be completed increases leading to an increase in patient in-service time.

We test for three Speedup mechanisms and find support for two. We find that nurses rush and spend less time in patient rooms providing treatment as the waiting room census increases. This is seen as a reduction in both the number of nurse-patient interactions and the mean duration of each interaction. We do not find evidence of doctors or nurses reducing the number of diagnostic tests ordered in response to crowding.

We introduce and find evidence for early task initiation. In the ED, diagnostic testing can be ordered early on by a triage nurse, or during the In-Service Phase by the physician. We find that when the waiting room is crowded and waiting times are long, triage nurses greatly increase the number of diagnostic tests ordered. This allows testing to be performed and processed while the patient is in the waiting room and thus reduces the required in-service time. This effect is most beneficial when the ED is crowded. However, we show that early task initiation leads to some amount of overtesting due to the nurses being unable to perfectly predict which tests will be desired by the doctor.

An interesting opportunity for future work is to develop a queuing model to identify the optimal level of early task initiation. (Baiman et al. (2010) models a queue with "helping resources" whish is conceptually realted to early task initiation.) Such a model must balance the service-time improvement benefits with the costs of overtesting (financial, medical, system congestion). In our setting, given the effectiveness of early task initiation in speeding up service and the fairly minimal overtesting potential (at least for abdominal pain patients), it is a bit surprising that diagnostic tests are not ordered in triage more regularly, regardless of the crowd level. Hospital managers should seek to understand the costs of overtesting and to find ways to minimize overtesting by improving triage nurses' ability to predict what tests will and won't be needed by the doctor. Redcuing the likelihood of overtesing would allow for early task initiation to be used even during times of only moderate crowding.

While we do not find evidence of task reduction, it is an operational lever that doctors and managers should at least consider. In the healthcare setting, task reduction is clearly a double-edged sword. On one hand, reducing testing speeds up service, reduces the load on the auxiliary services, and reduces costs. On the other hand, reduced testing may result in decreased quality of care. (While quality is not a focus of this study, we find no evidence of crowding leading to an increase in 72-hour revisits, a common ED quality metric.) In our discussions with several emergency medicine physicians we have heard that many tests are ordered either out of ritual or for "covering the bases" even though the probability of learning anything from the result is considered quite low. These types of tests should be identified and written into load-based care protocols that allow for skipping these tests when the ED is crowded.

Determining the "optimal" level of task reduction is an empirical medical question and is beyond the scope of this paper. Further, it is related to the philosophical question of what should be the role of the ED in the larger health care delivery system? Should the ED be the site of definitive medical care, or should it only serve to stabilize and route to the appropriate resource for full identification and care of the presenting medical condition? This is an ongoing debate in the medical community (Schuur and Venkatesh 2012, Wiler et al. 2012).

Lastly, we find that ignoring state-dependencies leads to inaccurate planning models. In our setting, the error was an overestimation of system busyness. Our results show that it is important to incorporate state-dependent mechanisms into planning models to avoid overinvestment in staffing and physical resources. Our results also show the value of identifying and measuring state-dependencies.

In conclusion, our work expands upon the prior state-dependent service time literature and shows that there can be several server-level mechanisms at work as servers respond to workload. We hope that incorporation of these mechanisms into future normative models will lead to better understanding and management of similar service systems with shared resources.

Appendix A: Log-Likelihood Function of Zero Inflated Negative Binomial Model

The zero inflated negative binomial model is estimated by maximization of the log-likelihood function. The function is derived from the combination of a logit model and a negative binomial count model. The function is given below and is based on the function shown in Hilbe (2011, p372). However, the formula in the book contains errors.

$$\mathcal{L}(\beta_1,\beta_2;y,\alpha)=\begin{cases}\ln\left(\frac{1}{1+\exp\left(-x_i'\beta_1\right)}\right)+\left(\frac{1}{1+\exp\left(x_i'\beta_1\right)}\right)\left(\frac{1}{1+\exp\left(x_i'\beta_2\right)}\right)^{1/\alpha} & \text{if } y=0\\ \ln\left(\frac{1}{1+\exp\left(x_i'\beta_1\right)}\right)+\frac{1}{\alpha}\ln\left(\frac{1}{1+\alpha\exp\left(x_i'\beta_2\right)}\right) & \\ +\ln\Gamma\frac{(y_i+1/\alpha)}{(y_i+1)(1/\alpha)}+y_i\ln\left(1-\frac{1}{1+\alpha\exp\left(x_i'\beta_2\right)}\right) & \text{if } y>0\end{cases}$$

References

Aksin, O. Zeynep, Patrick T. Harker. 2001. Modeling a phone center: Analysis of a multichannel, multiresource processor shared loss system. *Management Science* **47**(2) 324–336. doi:10.1287/mnsc.47.2.324.9842.

Alizamir, Saed, Francis de Véricourt, Peng Sun. 2013. Diagnostic accuracy under congestion. *Management Science* **59**(1) 157–171.

Armony, Mor, Shlomo Israelit, Avishai Mandelbarum, Yarvin N Marmor, Yulia Tseytlin, Galit B. Yom-Tov. 2013. On patient flow in hospitals: A data-based queuing-science perspective. *Working Paper* .

Baiman, Stanley, Serguei Netessine, Richard Saouma. 2010. Informativeness, incentive compensation, and the choice of inventory buffer. *The Accounting Review* **85**(6) 1839–1860.

Batt, Robert J., Christian Terwiesch. 2014. Waiting patiently: An empirical study of queue abandonment in an emergency department. *Working Paper* .

Berry Jaeker, Jillian, Anita L. Tucker. 2013. An empirical study of the spillover effects of workload on patient length of stay. *Working Paper* .

Bratti, Massimiliano, Alfonso Miranda. 2011. Endogenous treatment effects for count data models with endogenous participation or sample selection. *Health Economics* **20**(9) 1090–1109.

Caldwell, John A. 2001. The impact of fatigue in air medical and other types of operations: A review of fatigue facts and potential countermeasures. *Air Medical Journal* **20**(1) 25 – 32. doi:10.1016/S1067-991X(01)70076-4.

Centers for Medicare & Medicaid Services. 2012. Hospital outpatient prospective and ambulatory surgical center payment systems and quality reporting programs; electronic reporting pilot; inpatient rehabilitation facilities quality reporting program; quality improvement organization regulations. *Federal Register* **77**(146) 45061–45233.

Chan, Carri, Vivek F. Farias, Nicholas Bambos, Gabriel J. Escobar. 2011. Optimizing icu discharge decisions with patient readmissions. *Working Paper* .

Chen, Chao, Zhanfeng Jia, P. Varaiya. 2001. Causes and cures of highway congestion. *Control Systems, IEEE* **21**(6) 26 –32. doi:10.1109/37.969132.

Chisholm, Carey D, Edgar K Collison, David R Nelson, William H Cordell. 2000. Emergency department workplace interruptions are emergency physicians âĂIJinterrupt-drivenâĂİ and âĂIJmultitaskingâĂİ? *Academic Emergency Medicine* **7**(11) 1239–1243.

Crabill, Thomas B. 1972. Optimal control of a service facility with variable exponential service times and constant arrival rate. *Management Science* **18**(9) 560–566.

de Ven, Wynand P.M.M. Van, Bernard M.S. Van Praag. 1981. The demand for deductibles in private health insurance: A probit model with sample selection. *Journal of Econometrics* **17**(2) 229 – 252. doi:10.1016/0304-4076(81)90028-2.

Debo, Laurens G., L. Beril Toktay, Luk N. Van Wassenhove. 2008. Queuing for expert services. *Management Science* **54**(8) 1497–1512.

Dobson, Gregory, Tolga Tezcan, Vera Tilson. 2013. Optimal workflow decisions for investigators in systems with interruptions. *Management Science* **59**(5) 1125–1141.

Durbin, James. 1970. Testing for serial correlation in least-squares regression when some of the regressors are lagged dependent variables. *Econometrica: Journal of the Econometric Society* 410–421.

Fee, Christopher, Ellen J. Weber, Carley A. Maak, Peter Bacchetti. 2007. Effect of emergency department crowding on time to antibiotics in patients admitted with community-acquired pneumonia. *Annals of Emergency Medicine* **50**(5) 501–509.e1.

George, Jennifer M., J. Michael Harrison. 2001. Dynamic control of a queue with adjustable service rate. *Operations research* **49**(5) 720–731.

Gerla, M., L. Kleinrock. 1980. Flow control: A comparative survey. *Communications, IEEE Transactions on* **28**(4) 553 – 574. doi:10.1109/TCOM.1980.1094691.

Gilboy, N, T Tanabe, D Travers, AM Rosenau. 2011. *Emergency Severity Index (ESI): A Triage Tool for Emergency Department Care, Implementation Handbook*. Agency for Healthcare Research and Quality, Rockville, MD, 4th ed. AHRQ Publication No. 12-0014.

Greene, William H. 2012. *Econometric Analysis*. 7th ed. Prentice Hall.

Hilbe, Joseph M. 2011. *Negative Binomial Regression*. 2nd ed. Cambridge University Press.

Hopp, Wallace J., Seyed M. R. Iravani, Gigi Y. Yuen. 2007. Operations systems with discretionary task completion. *Management Science* **53**(1) 61–77.

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

Kc, Diwas Sign. 2013. Does multitasking improve performance? Evidence from the emergency department. *Forthcoming in Manufacturing and Services Operations Management* .

Kuntz, Ludwig, Roman Mennicken, Stefan Scholtes. 2014. Stress on the ward: evidence of safety tipping points in hospitals. *forthcoming in Management Science* .

Law, Averill M. 2007. *Simulation Modeling and Analysis*. 4th ed. McGraw Hill.

Little, John D. C. 1961. A proof for the queuing formula: L = ÃÕÂżw. *Operations Research* **9**(3) 383–387. doi:10.1287/opre.9.3.383. URL `http://pubsonline.informs.org/doi/abs/10.1287/opre.9.3.383`.

Lucas, Ray, Heather Farley, Joseph Twanmoh, Andrej Urumov, Nils Olsen, Bruce Evans, Hamed Kabiri. 2009. Emergency department patient flow: The influence of hospital census variables on emergency department length of stay. *Academic Emergency Medicine* **16**(7) 597–602.

McCarthy, Melissa L., Scott L. Zeger, Ru Ding, Scott R. Levin, Jeffrey S. Desmond, Jennifer Lee, Dominik Aronsky. 2009. Crowding delays treatment and lengthens emergency department length of stay, even among high-acuity patients. *Annals of Emergency Medicine* **54**(4) 492–503.e4.

Oliva, Rogelio, John D. Sterman. 2001. Cutting corners and working overtime: Quality erosion in the service industry. *Management Science* **47**(7) 894–914. doi:10.1287/mnsc.47.7.894.9807.

Pines, Jesse M., Judd E. Hollander. 2008. Emergency department crowding is associated with poor care for patients with severe pain. *Annals of Emergency Medicine* **51**(1) 1–5.

Pines, Jesse M., Judd E. Hollander, A. Russell Localio, Joshua P. Metlay. 2006. The association between emergency department crowding and hospital performance on antibiotic timing for pneumonia and percutaneous intervention for myocardial infarction. *Academic Emergency Medicine* **13**(8) 873–878.

Pines, Jesse M., Sanjay Iyer, Maureen Disbot, Judd E. Hollander, Frances S. Shofer, Elizabeth M. Datner. 2008. The effect of emergency department crowding on patient satisfaction for admitted patients. *Academic Emergency Medicine* **15**(9) 825–831.

Pines, Jesse M., Anjeli Prabhu, Joshua A. Hilton, Judd E. Hollander, Elizabeth M. Datner. 2010. The effect of emergency department crowding on length of stay and medication treatment times in discharged patients with acute asthma. *Academic Emergency Medicine* **17**(8) 834–839.

Schultz, Kenneth L., David C. Juran, John W. Boudreau, John O. McClain, L. Joseph Thomas. 1998. Modeling and worker motivation in jit production systems. *Management Science* **44**(12-Part-1) 1595–1607. doi:10.1287/mnsc.44.12.1595.

Schuur, Jeremiah D., Arjun K. Venkatesh. 2012. The growing role of emergency departments in hospital admissions. *New England Journal of Medicine* **367**(5) 391–393. doi:10.1056/NEJMp1204431.

Setyawati, L. 1995. Relation between feelings of fatigue, reaction time and work productivity. *Journal of Human Ergology* **24**(1) 129–135.

Staats, Bradley R., Francesca Gino. 2012. Specialization and variety in repetitive tasks: Evidence from a Japanese bank. *Management Science* **58**(6) 1141–1159. doi:10.1287/mnsc.1110.1482.

Stidham, Shaler, Richard R. Weber. 1989. Monotonic and insensitive optimal policies for control of queues with undiscounted costs. *Operations research* **37**(4) 611–625.

Tan, Tom, Serguei Netessine. 2012. When does the devil make work? An empirical study of the impact of workload on worker productivity. *Working Paper* .

Wiler, Jennifer L., Dennis Beck, Brent R. Asplin, Michael Granovsky, John Moorhead, Randy Pilgrim, Jeremiah D. Schuur. 2012. Episodes of care: Is emergency medicine ready? *Annals of Emergency Medicine* **59**(5) 351 – 357. doi:10.1016/j.annemergmed.2011.08.020.

Wolff, Ronald W. 1989. *Stochastic Modeling and the Theory of Queues*. Industrial and Systems Engineering, Prentice Hall Inc., Upper Saddle River, NJ.

Wooldridge, Jeffery M. 2009. *Introductory Econometrics: A Modern Approach*. 4th ed. South-Western Cengage Learning.

Yamazaki, Genji, Hirotaka Sakasegawa. 1987. An optimal design problem for limited processor sharing systems. *Management Science* **33**(8) 1010–1019. doi:10.1287/mnsc.33.8.1010.